Journal of Occupational and Organizational Psychology (2007), 80, 575–600
© 2007 The British Psychological Society

The British Psychological Society
www.bpsjournals.co.uk

Measurement equivalence in the conduct of a global organizational survey across countries in six cultural regions

Alain De Beuckelaer[1]*, Filip Lievens[2] and Gilbert Swinnen[3]
[1]Radboud University Nijmegen, The Netherlands
[2]Ghent University, Belgium
[3]Hasselt University, Belgium

This study examined the measurement equivalence of a global organizational survey measuring six work climate factors as administered across 25 countries ($N = 31,315$) in all regions of the world (West Europe, East Europe, North America, Latin America, South America, Middle East, Africa and Asia-Pacific). Across all countries, the survey instrument exhibited 'form equivalence' and 'metric equivalence', suggesting that respondents completed the survey using the same frame-of-reference and interpreted the rating scale intervals similarly. Schwartz's (1994, 1999, 2004) cultural value theory was then used for grouping the countries in cultural regions, and to anticipate measurement equivalence of the data from the survey within and between these regions. Results showed partial support for Schwartz's theory. The English-speaking region was the only region where empirical evidence for 'scalar equivalence' was found. No support was found for the prediction that measurement equivalence would be higher among countries that are part of cultural regions with a small cultural distance than among countries that are part of cultural regions with a large cultural distance. However, the use of a common language in a particular cultural region reduced the bias present in the cross-country comparison within that region.

When psychological and work-related constructs are measured in a cross-cultural context, it is pivotal to establish equivalence of the measures prior to drawing meaningful substantive conclusions about the relative importance of constructs across countries (Little, 1997; Riordan & Vandenberg, 1994; Schaffer & Riordan, 2003; Vandenberg & Lance, 2000; Van de Vijver & Leung, 1997). Lack of measurement equivalence in data across countries implies that there is no common basis to compare data across countries: In such case, observed mean differences on relevant constructs (across countries) might result from measurement artefacts related to the measurement instrument used rather than from true differences across countries.

*Correspondence should be addressed to Dr Alain De Beuckelaer, Nijmegen School of Management, Radboud University Nijmegen, Thomas van Aquinostraat 1, Nijmegen, 6500 HK, The Netherlands (e-mail: A.DeBeuckelaer@fm.ru.nl).

DOI:10.1348/096317907X173421

Establishing measurement equivalence enables us to answer a series of important questions (see Table 1) such as: Do respondents in different countries use a similar frame-of-reference when answering items used to measure relevant constructs? Do respondents in different countries calibrate the intervals on the measurement scale used in similar ways? Are differences in response styles across countries (e.g. the tendency to say 'yes' or to use extreme response categories) partly responsible for observed cross-country differences in mean item scores?

The purpose of this study is to examine the measurement equivalence of a survey instrument[1] across 25 countries. As compared to the majority of measurement equivalence studies, a key contribution of this study is that we test and report whether Schwartz's (1994, 1999, 2004) theory of cultural values can be used as a theoretical framework for explaining why measurement equivalence can/cannot be established across countries and cultural regions. Our study is a truly global endeavour as we examine measurement equivalence across countries in West Europe, East Europe, North America, Latin America, South America, Middle East, Africa and Asia-Pacific. Although other culture frameworks exist (Hofstede, 1980, 1991; Kluckhohn & Strodtbeck, 1961; Trompenaars & Hampden-Turner, 1998), Schwartz's theory is especially useful for our study because data on countries' culture value priorities are available from countries spread across every continent in the world.

This study is situated in the context of international surveying (Johnson, 1996; Ryan, Chan, Ployhart, & Slade, 1999; Saari & Judge, 2004). One of the most surveyed topics in organizational surveys is employees' climate perceptions (Church & Waclawski, 2001; Kraut, 1996; Rucci, Kirn, & Quinn, 1998). In recent years, climate surveys have become increasingly popular as tools for organizational diagnosis and change because employee perceptions of climate have been found to be linked to important individual and organizational outcomes (Parker *et al.*, 2003). Generally, climate surveys seem to capture five primary domains of the work environment, namely job characteristics, role characteristics, leadership characteristics, work-group/social characteristics and organizational characteristics (James & Jones, 1974; James & McIntyre, 1996; James & Sells, 1981; Parker *et al.*, 2003). Climate surveys can be conducted using either a standard instrument (i.e. existing, established) or a customized (i.e. organization-specific) instrument, with each having its advantages and disadvantages (Church & Waclawski, 2001, p. 52). The instrument under investigation in this study is a customized instrument because only a customized climate survey provided sufficient in-depth information needed for achieving the organization's specific set of objectives (Church & Waclawski, 2001, p. 52). Along these lines, Kraut (2006) posits that, in order to be effective, organizational surveys have to be tied to the (desired) organizational outcomes (e.g. specific action points), implying that they should be customized to a large extent. Despite being a customized instrument, the instrument under investigation included measures of four of the five domains of climate surveys (see Parker *et al.*, 2003), attesting to its generalizabililty to other standard and/or customized climate surveys.

[1] *Measurement equivalence is not a property of the survey instrument itself, but a property of a particular administration of the survey instrument. For the reader's convenience we will refer in this paper to 'measurement equivalence of the survey instrument' when we actually mean 'measurement equivalence of the instrument as administered in the particular survey under study'. Furthermore, by using the term 'survey instrument' we do not refer to all survey questions, but only those that actually measure the constructs that are of interest to the study.*

Table 1. Measurement equivalence tests and their conceptual meanings

Measurement equivalence model	Statistical test	Conceptual meaning	Implications
(Factor) Form equivalence model	Equivalent pattern of salient and non-salient factor loadings across countries.[1] To set a metric for the factor, the loading of one indicator per factor (i.e. the *reference* indicator) is constrained to one in all countries	There are no cross-country differences in respondents' frame-of-reference when completing the instrument	All factors are measured by an identical set of indicators in all countries
Metric equivalence model	All factor loadings are constrained to be identical across countries	There are no cross-country differences in respondents' calibration of the intervals on the measurement scale. Differences in Extreme Response Style (ERS) across countries are not likely	Structure-level comparisons (i.e. dealing with cause-effect relationships) across countries are meaningful
Scalar equivalence model	All factor loadings and indicator intercepts are constrained to be identical across countries	Differences in Acquiescence Response Style (ARS) (i.e. agreement bias) across countries are not likely	Structure-level comparisons and level-oriented comparisons (e.g. based on estimated construct means) across countries are meaningful

Note. [1] In the (factor) form equivalence model tested in this study all indicators load on one factor only (i.e. double-loadings are not allowed).

Study background

Prior measurement equivalence research of survey instruments

Riordan and Vandenberg (1994) posited that people's general values influence their work-related goals and values and that, therefore, these values serve as a frame-of-reference against which they define their work-related experiences. As surveys ask for reports on work-related experiences, it can be assumed that individuals with different values will not always use the same frame-of-reference when completing survey items. For instance, an organizational concept such as privacy might have a different conceptual meaning across cultures. In cultures wherein individuals are strongly embedded in social groups, items related to privacy might be interpreted differently than in cultures wherein individuals are more autonomous (see Hulin & Mayer, 1986). Clearly, such differences in the conceptual domain for interpreting survey items might decrease the measurement equivalence of global surveys. Conversely, similarity in cultural values might increase the use of similar conceptual domains when completing global surveys, resulting in measurement equivalence of the scales used in the survey.

Apart from affecting the conceptual frame-of-reference used, cultural values might also influence how individuals interpret the rating scale (Riordan & Vandenberg, 1994). Specifically, prior cross-cultural research has shown that the differences between the intervals of a rating scale are differently perceived across cultures. In fact, substantial cross-country differences have been found with regard to the tendency to agree with items, regardless of the item content (Cunningham, Cunningham, & Green, 1977; Grimm & Church, 1999; Morris & Pavett, 1992; Riordan & Vandenberg, 1994; Ross & Mirowski, 1984; Van Herk, Poortinga, & Verhallen, 2004). Similarly, there is empirical evidence for cross-country bias due to the respondents' use of extreme responses on rating scales as this bias exists between Korean and American respondents (e.g. Chun, Campbell, & Yoo, 1974; Lee & Green, 1991), Japanese and American respondents (e.g. Stening & Everett, 1984; Zax & Takabashi, 1967), and French and Australian respondents (Clarke, 2000). Such cross-country differences in response styles (see also Johnson, Kulesa, Cho, Young, & Shavitt, 2005) produce systematic differences in observed variable means and variances. As a result, the assumption of measurement equivalence of survey instruments may not be tenable. Although these prior studies were not conducted in an organizational (survey) context, they might have direct implications for organizational surveys because the latter also use rating scales.

Some recent research (Liu, Borg, & Spector, 2004; Ryan *et al.*, 1999) has tackled similar questions with regard to the cross-cultural equivalence of organizational surveys across multiple countries. Ryan *et al.* scrutinized the equivalence of an organizational survey of a multinational company across four countries (Mexico, U.S., Australia and Spain). They found that the organizational survey was equivalent across U.S. and Australian samples only. Recently, Liu *et al.* examined whether the German Job Satisfaction Survey was 'transposable' across 18 countries. These countries were located in four cultural regions of Schwarz's (1994, 1999, 2004)[2] cultural model, namely West Europe, Far East, English-speaking region and South America (i.e. Latin cultural region). Two other regions (East Europe and Islamic countries) were not included in their study. Liu *et al.* concluded that the German Job Satisfaction Survey was equivalent only across

[2] *Strictly speaking, Schwartz' latest coplot analysis (Schwartz, 2004) also includes a seventh cultural region (Confucian-influenced countries, such as China, Japan, South Korea and Vietnam). In our analyses, this cultural region is integrated in the Far Eastern cultural region (as in Schwartz, 1999).*

countries sharing the same cultural background and language. For example, measurement equivalence was established across countries within the same cultural region. In addition, the satisfaction survey was more equivalent among countries in similar cultural regions than among countries in distant cultural regions.

This study's measurement equivalence hypotheses

In this study, we used Schwartz's (1994, 1999, 2004) model of cultural values as a theoretical framework for explaining why measurement equivalence can/cannot be established across countries and cultural regions (see also Liu *et al.*, 2004). Similar to other culture frameworks (e.g. Hofstede, 1980, 1991; Kluckhohn & Strodtbeck, 1961; Trompenaars & Hampden-Turner, 1998), Schwartz (1994, 1999) posited that cultural value dimensions represent the key issues that all societies face. Specifically, seven types of values are distinguished. These seven value types are organized along three polar dimensions. The first dimension refers to the relation between the individual and the group. According to Schwartz (1999), there are two opposing ways of resolving this basic issue. On one hand, there are cultures described by Conservatism (also referred to as Embeddedness). In such cultures, the *status quo* is maintained by embedding individuals in groups. Individuals find meaning only as part of the social order and social relationships. On the other hand, there are cultures characterized by Autonomy. In these cultures, individuals are encouraged to independently and voluntarily pursue their own ideas (i.e. Intellectual Autonomy) and emotions (i.e. Affective Autonomy). The second dimension deals with how cultures ensure socially responsible behaviour. Again, two opposing ways of resolving this issue are distinguished: Hierarchy vs. Egalitarianism. Cultures described by Egalitarianism socialize people to voluntarily cooperate with others and to be genuinely concerned about others. Cultures described by Hierarchy socialize people in such a way as to make them follow the rules attached to their roles. So, there exists an unequal power distribution. The third dimension describes how societies relate to their social and natural environment. At one pole of this dimension ('Mastery'), people try to take charge, change their environment and utilize it. At the other pole of this dimension one finds 'Harmony'. This value type relates to accepting the world as it is and trying to fit in it rather than to modify it.

Schwartz's theory has been validated with over 35,000 respondents from 122 samples in 49 nations from every continent in the world. Accordingly, it is possible to compare cultures on the basis of their emphases on the seven types of values. Statistical analyses consistently revealed almost identical mappings of world cultures (see Schwartz, 1994, 1999, 2004; Schwartz & Bardi, 1997; Schwartz & Ros, 1995). Specifically, on the basis of their cultural value priorities countries could be meaningful grouped in 6 broad cultural regions: Western European countries (characterized by Intellectual Autonomy, Egalitarianism and Harmony), English-speaking countries (characterized by Mastery and Affective Autonomy), Far Eastern countries (characterized by Hierarchy and Conservatism), East European countries (characterized by Conservatism and Harmony), Latin American countries (characterized by moderate levels of all seven value types) and Islamic countries (characterized by moderate levels on Conservatism and Hierarchy). According to Schwartz (1999), these broad meaningful groupings of countries are not only based on shared valued priorities but also on geographical proximity, shared histories, religion, level of development, culture contact and other factors (see also Schwartz & Bardi, 1997; Schwartz & Ros, 1995).

In light of Schwartz's theory and research, we hypothesize that measurement equivalence of the organizational survey will be established across countries that are part of the same cultural region. Hence, we hypothesize measurement equivalence of the instrument across countries *within each of the six cultural regions*: West Europe, English-speaking, Far East, East Europe, Latin America and Islamic countries. Although prior studies examined this hypothesis across a limited number of countries within specific cultural regions (Liu *et al.*, 2004; Schwarzer *et al.*, 1997; Spector *et al.*, 2002), no study has tested this hypothesis in all six cultural regions of the world.

A second set of hypotheses deals with between-culture measurement equivalence. This second set of hypotheses is based on the notion of cultural distance. Cultural distance between national cultures can be conceptualized as a dissimilarity or distance measure between two countries' scores on key cultural dimensions (Kogut & Singh, 1988; Manev & Stevenson, 2001; Shenkar, 2001). As explained before, the key cultural dimensions considered in this study comprise the seven country value types as identified by Schwarz (Schwartz, 1994, 1999, 2004). As our second set of hypotheses deals with measurement equivalence between cultural regions, our emphasis is mainly on making direct comparisons between different cultural regions, excluding comparisons between countries belonging to the same cultural region. In line with our first set of hypotheses, countries belonging to the same cultural region are supposed to share similar values, and thus not really culturally distant. A similar approach to making comparisons between cultural regions was adopted by Liu *et al.* (2004). Do notice that cultural distance between different cultural regions is visually depicted in Schwarz's (1999, 2004) coplot analysis. In fact, if one looks at how the cultural regions are positioned in Schwartz's coplot analysis (see Schwarz, 2004, p. 58), it becomes clear that some cultural regions are more close to each other than other cultures. For example, the cultural regions of West Europe and English-speaking cultural regions are located next to each other in Schwartz's (1999, 2004) coplot results. This means that the value emphases (and other factors) of these cultural regions are more similar than other cultural regions. Hence, we expect that individuals in countries in these cultural regions will have a more similar frame-of-reference when completing the survey and will calibrate the intervals of the rating scale more similarly than individuals in countries in cultural regions with a large cultural distance, leading to measurement equivalence.

As an example of cultural regions with a larger cultural distance, East Europe is positioned oppositely to the English-speaking cultural region. As noted above, English-speaking countries are described by Mastery and Affective Autonomy values at the expense of Conservatism and Harmony values, whereas the opposite is the case in the East European countries. As another example of cultural regions with a larger cultural distance, West Europe is located at the other side of the Far East in terms of the Intellectual Autonomy, Egalitarianism and Harmony. Finally, there is also a large distance between the Islamic cultural region and either the West European or English-speaking regions on the Autonomy vs. Conservatism dimension. Thus, in investigating measurement equivalence of survey results across countries that are part of cultural regions with a large cultural distance, we hypothesize that employees in these countries will use a different conceptual domain and will use the rating scale differently when completing the survey so that measurement equivalence will not be established.

Taken together, we formulate two set of hypotheses. One set of hypotheses deals with the establishment of measurement equivalence *within cultural regions*. A second set of *between - region* hypotheses states that measurement equivalence will be higher among countries that are part of cultural regions with a small cultural distance than among countries that are part of cultural regions with a large cultural distance.

Method

Sample and procedure

Individual-level data from 31,315 managers in 25 countries were collected within a multinational company in the fast moving consumer goods sector. The multinational has business operations in more than 70 countries across the world. Countries were selected for inclusion in this study only when the sample size exceeded 100. This minimum sample size was chosen because it resulted in a subject-variable ratio higher than 5 to 1 in the analyses. Table 2 presents the list of countries in this study broken down by the cultural regions of Schwartz.

In principle, all managers were being surveyed. In the United States, however, some lower-level managers were not surveyed (due to union restrictions). The same was true for African countries. The people surveyed were informed in advance about the purpose of the survey, the content coverage and the confidentiality of the data provided. The goal of the organizational survey was (1) to enhance employees' involvement and motivation and (2) to provide baseline data for organizational change efforts (i.e. towards the new business strategy). To increase employees' awareness, posters and short articles were posted on the intranet together with a letter from the CEO. Respondents individually completed web-based surveys[3] at their work site. Reminders were sent to people who had not responded two weeks after data collection started, and just before closing the survey administration phase. In total, data collection took about four months. The overall response rate across countries was 76.4%, which is above the average survey response rate given by Church and Waclawski (2001) and Kraut (1996). No response rates per country were made available to the authors.

Organizational survey

The organizational survey under investigation in this study was constructed analogously to corporate sponsored global surveys (see Johnson, 1996). This meant that the HR staff at the corporate headquarters led the development and administration of the survey. In addition, a common method and framework to survey employees across countries was followed, while allowing for country customization. The first step in the survey design process involved composing a global survey team. This global survey team consisted of: (1) a broad cross-section of employees from different levels, functional areas, and backgrounds of the multinational company and (2) survey consultants. Next, the global survey team developed the English items of the source questionnaire used for translation purposes. Some of the items had been used before by the survey consultants. Others were added by the newly constituted global survey team. The closed-ended items of the survey used 5-point Likert-type response formats. The predominant response format ranged from *strongly disagree* (1) to *strongly agree* (5). Other items were scored on a scale ranging from *very poor* (1) to *very good* (5). In the following step, regional survey leaders were responsible for the translations needed within their region. They supervised and monitored the different translations which were checked by local survey co-ordinators using the English survey as the basis for comparison. Next, professional interpreters back

[3] *Consistent with corporate-sponsored global surveys, a common methodology and framework was used, while allowing country customization. Therefore, in some countries paper-and-pencil survey administration had to be used instead of web-based administration. It seems unlikely that these different administration modes confound our examination because prior research has not revealed major threats to measurement equivalence across web-based and paper-and-pencil survey administration modes (Cole, Bedeian, & Field, 2006; Fenlason & Zuckow-Zimberg, 2006; Stanton, 1998).*

Table 2. List of study countries broken down by Schwartz's cultural region

Cultural region	Country	Language used	Males (%)	Expatriates (%)	Job level (low; intermediate; high) (%)	Managers working in production, maintenance, quality control and engineering (%) [+rank]
1. West Europe	Belgium (*N* = 668)	Dutch (65%)/French (35%)	64.1	4.0	78.8; 13.3; 7.9	22.8 [18]
	France (*N* = 829)	French	57.1	3.0	59.4; 30.0;10.6	34.1 [16]
	Germany (*N* = 1142)	German	70.9	2.5	70.5; 23.1; 6.4	35.6 [15]
	Italy (*N* = 1293)	Italian	75.5	1.4	76.5; 13.4; 10.4	44.7 [6]
	Netherlands (*N* = 921)	Dutch (90%)/English (10%)	64.7	11.0	58.4; 28.5; 13.1	16.0 [19]
	Sweden (*N* = 517)	Swedish	62.9	1.6	84.5; 11.7; 3.8	41.2 [9]
	Switzerland (*N* = 235)	German (85%)/French (15%)	64.9	5.4	50.4; 37.0; 12.6	5.4 [25]
2. East Europe	Hungary (*N* = 729)	Hungarian	56.1	1.0	86.0; 9.9; 4.1	42.7 [8]
	Russian Federation (*N* = 695)	Russian	47.9	5.9	89.2; 8.2; 2.6	54.6 [1]
3. English-speaking	Australia (*N* = 951)	English	53.1	0.8	81.6; 12.7; 5.7	40.6 [11]
	Canada (*N* = 776)	English (95%)/French (5%)	48.6	2.1	77.3; 16.6; 6.1	25.1 [17]
	South Africa (*N* = 651)	English	71.8	3.3	86.3; 9.5; 4.2	50.8 [3]
	UK (*N* = 2728)	English	55.2	5.0	70.0; 21.3; 8.7	38.8 [12]
	US (*N* = 4694)	English	54.5	2.4	70.7; 21.3; 8.0	35.8 [14]
4. Latin/South America	Argentina (*N* = 2278)	Spanish	78.2	0.8	90.2; 6.6; 3.2	40.7 [10]
	Brazil (*N* = 6421)	Portuguese	80.4	1.6	90.3; 4.8; 4.9	51.6 [2]
	Chile (*N* = 919)	Spanish	86.8	0.6	86.9; 8.9; 4.2	50.0 [4]
	Honduras (*N* = 449)	Spanish	62.2	0.4	94.6; 1.7; 3.7	43.3 [7]
	Mexico (*N* = 1901)	Spanish	77.1	0.9	89.8; 6.6; 3.6	47.7 [5]
5. Far East	China (*N* = 150)	Mandarin	60.4	17.4	15.3; 70.2; 15.3	15.5 [21]
	Japan (*N* = 260)	Japanese	86.3	6.8	43.1; 42.7; 14.2	9.9 [22]
6. Islamic countries	Egypt (*N* = 313)	Arabic	79.3	2.6	82.5; 12.1; 5.4	36.3 [13]
	Malaysia (*N* = 118)	English (100%)	59.9	2.1	25.2; 60.1; 14.7	7.0 [23]
	Pakistan (*N* = 590)	English (75%)/Urdu (25%)	90.9	1.7	76.1; 16.6; 7.3	15.7 [20]
	Turkey (*N* = 1087)	Turkish	85.9	1.1	84.8; 11.3; 3.9	5.6 [24]

Note. All percentages on background variables (% males, % expatriates etc.) are based on individual-level information. There is, however, one exception. As far as the variable 'language used' is concerned, no individual-level information was made available. Hence, the percentages reported in this column are based on aggregate level data.

translated the surveys to English. Finally, English-speaking Master students in Management compared the back translation to the original English version and indicated whether the meaning of each item had remained similar. Generally, results were satisfactory. The final surveys were pilot tested in each individual country and -if necessary- modifications were made.

The final survey may be broken into three parts, which together consisted of 102 items. One part asked for organizational members' reports and perceptions of this multinational's core dimensions. Specifically, the following six work environment factors were considered to be of key importance for the multinational: team commitment, supervisor support, goal clarity, decision-making, organizational adaptability and environmental and societal responsibility. The Appendix presents the definitions of these work environment factors. The second part of the survey dealt with background information (e.g. age, gender, tenure), whereas the third part comprised country-specific questions. In this study, we focused on a specific set of items (15) in this global survey. Consistent with Ryan *et al.* (1999), we retained only items that were 'clearly'[4] linked to the six work environment factors. The Appendix presents these items.

Although the six constructs measured in this organizational survey were specific to this company, we emphasize that they correspond well to the constructs typically included in organizational climate surveys (James & Jones, 1974; James & McIntyre, 1996; James & Sells, 1981; Parker *et al.*, 2003). In fact, the organizational survey under investigation measured four of the five core organizational climate dimensions, namely work-group/social characteristics (i.e. team commitment), leadership characteristics (i.e. supervisor support and decision-making), role characteristics (i.e. goal clarity) and organizational characteristics (i.e. organizational adaptability and environmental and societal responsibility). Only job characteristics were not measured in this organizational survey.

Analyses

Overview

To test our hypotheses, three types of analyses were conducted. First, we examined measurement equivalence across all countries simultaneously. Second, we conducted within-region measurement equivalence analyses. Third, we conducted between-region measurement equivalence analyses. In all these analyses, the same set of models was tested. The remainder of the paper discusses these various measurement equivalence models. All statistical analyses conducted made use of individual-level information (i.e. every person's responses to the survey questions).

Sequence of models tested

We started by testing a confirmatory factor analysis model which imposed the hypothesized six-factor structure (i.e. the six key factors of the organization's business

[4] We first screened the 102 survey items in terms of content/item format. Accordingly, we removed 9 items that measured job satisfaction, 12 items that used a response format other than a Likert-scale (e.g. a binary response format), and 53 items that were not tied to a specific construct (i.e. one-item measures). Second, we statistically screened the items because CFA is a restrictive statistical technique that puts very strong demands on the psychometric properties of the items used to operationalize the construct. Therefore, we conducted an exploratory factor analysis on the remaining 28 items as precursor to the CFA (Gerbing & Hamilton, 1996; Hurley et al.*, 1997). This lead to the further removal of 13-items. This leaves us with 15 items measuring the six work environment factors considered in the study. A split-half analysis (half of the sample used for factor-analysis and half of the sample used for testing the derived factor structure) confirmed the 6-factor structure.*

model) onto the data. Strictly speaking, this model is not a mean- and covariance structure (i.e. MACS) model as indicator mean scores are not needed to test the hypothesized dimensionality of the factor model. The six-factor model was evaluated using the samples from all individual countries.

Provided that the data fit the six-factor model well, a hierarchical sequence of nested statistical models (e.g. Vandenberg & Lance, 2000) can be used to assess measurement equivalence of indicator variables across countries (see Table 1). In the methodological literature, there is some debate as to which minimal set of measurement parameters should be identical across groups (countries). In this study, we investigated whether Meredith's (1993) (relatively) strong definition of measurement equivalence would be realistic for the data. According to Meredith, factor loadings and indicator intercepts of observed variables should be identical across groups (countries). Unique variances of indicators (i.e. unreliabilities) may, however, vary across countries. The same is true for factor means, factor variances and factor covariances. Meredith's equivalence condition is referred to as 'scalar equivalence' across groups. According to Meredith (1993), Little (1997) and Chan (2000), scalar equivalence across groups is a prerequisite for the comparison of (latent) factor means.

Taking Meredith's (1993) scalar equivalence model as measurement equivalence criterion, we conducted the following set of increasingly restrictive tests of measurement equivalence. First, we specified a baseline model in which no parameters (i.e. factor loadings, indicator intercepts, unique variances, factor means, and factor variances and covariances), except for the factor loading of the reference indicator, which were constrained to be equal to one in all countries (see Table 1). Conceptually, the baseline model assumes that the data exhibit (factor) 'form equivalence' across countries. In other words, the observed variables are assumed to be related to the same number of factors and the factors are measured by the same set of observed variables in all countries. An additional constraint in our baseline model is that all observed variables load on just one factor (i.e. cross-loadings are not specified). If form equivalence across countries is established, factor structures underlying employees' ratings are the same in the countries under study. Conceptually, this means that employees use a similar frame-of-reference when completing the items of the organizational survey (Riordan & Vandenberg, 1994; Vandenberg & Lance, 2000).

The second model in the sequence constrains all factor loadings to be identical across countries while all other parameters (i.e. indicator intercepts, unique variances, factor means and factor variances and covariances) are freely estimated. This model is called the 'metric equivalence model'. The model of metric equivalence assumes that the factor loadings of all observed variables are identical across countries. Conceptually, equivalence of factor loadings implies that respondents calibrate the intervals used on the measurement scale in similar ways (Riordan & Vandenberg, 1994; Vandenberg, 2002). This makes it possible to draw meaningful structure-level comparisons across countries (i.e. comparisons dealing with cause-effect relationships). Non-equivalence of factor loadings (across countries) would imply that there are substantial differences across countries in terms of the extent to which observed variable scores change as a result of a fixed (e.g. a unit) change in the underlying factor score(s). This may stem from cross-country differences in the use of extreme response style (ERS, see also Cheung & Rensvold, 2000). This is because extreme responses might produce a larger variance on the observed variables for a stronger ERS country than for a weaker ERS country. In-turn, cross-country differences in the variance structure of observed variables might lead to substantial cross-country differences in terms of the factor loading of observed variables.

The third model in the sequence, the scalar equivalence model, constrains all factor loadings and indicator intercepts to be identical across countries. The remaining parameters (i.e. unique variances, factor means and factor variances and covariances) are not constrained across countries. This model provides sufficient evidence to conclude that the measurement scale used to score the observed variables (i.e. the indicators of constructs) is identical across countries (Drasgow, 1984, 1987), enabling us to draw meaningful level-oriented comparisons across countries (i.e. comparisons based on estimated construct means). Non-equivalence of indicator intercepts across countries may be caused by cross-country differences in agreement bias (also known as 'acquiescence response style bias') (see Cheung & Rensvold, 2000). This is because a higher tendency to respond positively to items in one country (as opposed to all other countries) leads to a higher scale origin in that particular country. As a result, the estimate for the factor mean is inflated due to the additive bias caused by the higher indicator intercept in that country.

Assessment of model fit

To assess the fit of the models, we relied on three measures of model fit which are less sensitive to sample size than the traditional Chi-squared statistic. In particular, the following goodness–of-fit measures were used: (1) the Comparative Fit Index (Bentler, 1990), (2) the Tucker-Lewis Index (TLI), which is also referred to as the Bentler-Bonnet Non-Normed Fit Index (NNFI) (Bentler & Bonett, 1980), (3) the Root Mean Square Error of Approximation (RMSEA) (Steiger, 1990). These goodness–of-fit measures were suggested by Hu and Bentler (1999). Their extensive simulation study evaluated the adequacy of cut-off values based on the criterion that the adequate cut-off values should result in minimum type I and type II errors. On the basis of this study, Hu and Bentler proposed the following cut-off values: .95 (i.e. minimum values for CFI and TLI) and .06 (i.e. maximum value for RMSEA).

To statistically compare alternative measurement equivalence models (such as the form equivalence model, the metric and the scalar equivalence model), the Chi-squared difference statistic is traditionally used. However, the Chi-squared difference statistic is also sensitive to sample size (Brannick, 1995; Kelloway, 1995). Recently, a simulation study assessed the usefulness of many other measures of model fit (in addition to Chi-squared difference statistic) when statistically comparing alternative models specifying different levels of measurement equivalence across groups (Cheung & Rensvold, 2002). The difference in Comparative Fit Index (CFI) between nested equivalence models emerged as a more reliable (and robust) measure of model fit than the classical Chi-squared difference test. More specifically, the difference in CFI between (successive) equivalence models should not be higher than .01. The difference in CFI was, therefore, used to choose between alternative models varying in terms of the level of measurement equivalence assumed to be present in the data.

Results

Descriptive statistics

Table 3 provides an overview of the average scores of all countries on each of the six factors. The average scale values have been calculated for descriptive purposes only. Such values are useful to get a rough idea about score differences across countries, but

Table 3. Descriptive statistics per country on the six measures

	Team commitment (F1)		Supervisor support (F2)		Goal clarity (F3)		Decision making (F4)		Organizational adaptability (F5)		Environmental and societal responsibility (F6)	
Country	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Belgium	3.78	.81	3.73	1.02	4.24	.72	3.82	.90	3.01	.66	4.06	.80
France	3.70	.85	3.51	1.15	4.04	.87	3.75	.91	2.76	.70	4.12	.76
Germany	3.83	.78	3.51	1.13	4.06	.86	3.01	1.09	2.46	.70	3.95	.84
Italy	3.73	.88	3.81	1.04	3.85	.93	3.79	.98	3.38	.71	4.16	.85
Netherlands	3.80	.82	3.70	1.06	4.34	.69	3.76	.88	2.81	.68	4.22	.70
Sweden	3.71	.84	3.60	1.20	3.69	.96	3.55	1.15	2.57	.75	3.72	.97
Switzerland	4.01	.78	3.90	1.12	4.22	.77	3.67	1.00	2.68	.77	3.88	.87
Hungary	3.88	.77	4.02	.99	4.23	.76	4.08	.84	2.80	.69	4.12	.78
Russian federation	4.01	.82	4.02	.98	4.34	.76	4.04	.88	2.98	.68	4.48	.70
Australia	3.67	.87	3.72	1.13	4.13	.86	3.61	1.02	2.82	.74	4.27	.62
Canada	3.73	.95	3.72	1.16	4.00	.95	3.84	1.02	2.70	.76	4.35	.78
South Africa	3.75	.93	3.66	1.17	4.37	.76	3.97	.98	3.09	.74	4.47	.72
UK	3.54	.91	3.39	1.15	3.93	.92	3.43	1.08	2.63	.76	4.01	.84
US	3.72	.95	3.74	1.17	4.10	.89	3.78	1.07	2.79	.84	4.30	.81
Argentina	4.13	.79	3.91	1.17	4.50	.66	4.12	.92	3.47	.65	4.68	.56
Brazil	3.44	.99	3.67	1.22	4.18	.88	3.99	1.08	3.23	.83	4.45	.69
Chile	3.99	.85	3.58	1.22	4.21	.83	3.81	1.04	3.26	.69	4.63	.64
Honduras	4.13	.78	3.88	1.11	4.33	.81	3.89	1.10	3.49	.74	4.52	.76
Mexico	3.98	.88	3.74	1.19	4.41	.74	4.01	1.02	3.29	.74	4.63	.62
China	3.61	.92	3.82	1.02	4.34	.64	3.79	.82	2.59	.65	4.11	.83
Japan	3.72	.71	3.76	.93	4.27	.60	3.58	.81	2.65	.68	3.30	.98
Egypt	3.97	.82	3.98	1.07	4.10	.86	3.83	1.01	3.21	.73	4.34	.84
Malaysia	3.69	.81	3.88	1.01	4.24	.74	3.86	.84	2.57	.67	4.20	.72
Pakistan	3.92	.89	4.03	1.09	4.50	.67	4.10	.94	2.94	.81	4.56	.71
Turkey	4.22	.76	3.80	1.12	4.68	.53	4.28	.84	3.34	.73	4.81	.47

they fall short in terms of their measurement properties (e.g. the calculated average scale values capitalize on item unreliabilities [i.e. measurement error]).

Test of baseline six-factor CFA model

Confirmatory factor analyses were run to test the adequacy of the six-factor structure using the data from each individual country. As shown in Table 4, the six-factor model provided an adequate representation of the data in all countries, except for South Africa. As this model is a prerequisite for any subsequent measurement equivalence tests (see Netemeyer, Bearden, & Sharma, 2003, p. 156), we decided to restrict our further analyses to 24 countries only (i.e. excluding South Africa). In Japan, one of the four measures of model fit (i.e. TLI) did not satisfy the criteria of Hu and Bentler (1999). As all other measures of model fit report very acceptable values for Japan, the data from Japan were still used in subsequent analyses. In short, in virtually all countries, the six-factor model provided a satisfactory model fit. Therefore, we could continue with conducting the measurement equivalence tests.

Table 4. SEM Models specifying a six-factor structure

Country	χ^2 (*df*)	CFI	TLI	RMSEA
Belgium	189.9 (75)	.965	.951	.048
France	159.1 (75)	.981	.973	.037
Germany	264.1 (75)	.969	.956	.047
Italy	262.5 (75)	.974	.963	.044
Netherlands	201.8 (75)	.970	.957	.043
Sweden	181.5 (75)	.969	.956	.052
Switzerland	116.2 (75)	.970	.958	.048
Hungary	189.0 (75)	.969	.956	.046
Russian federation	176.4 (75)	.969	.956	.044
Australia	237.5 (75)	.974	.963	.048
Canada	212.1 (75)	.977	.968	.049
South Africa	275.1 (75)	.946	.924	.064
U.K.	479.4 (75)	.978	.969	.044
U.S.	1156.0 (75)	.968	.955	.055
Argentina	339.9 (75)	.979	.970	.039
Brazil	812.0 (75)	.976	.966	.039
Chile	200.3 (75)	.972	.961	.043
Honduras	117.3 (75)	.978	.970	.035
Mexico	366.4 (75)	.972	.960	.045
China	105.6 (75)	.967	.953	.052
Japan	125.7 (75)	.954	.935	.051
Egypt	103.3 (75)	.980	.973	.035
Malaysia	97.3 (75)	.976	.967	.050
Pakistan	185.1 (75)	.969	.956	.050
Turkey	194.1 (75)	.974	.963	.038

Overall measurement equivalence tests

We used structural equation modelling in the form of mean and covariance structure (MACS) analysis models to test for measurement equivalence. Mplus 2 (Muthén & Muthén, 1999, 2003) was used to evaluate all MACS models.

All measurement equivalence tests across all countries are shown in Table 5. From the goodness-of-fit statistics reported it is clear that the form equivalence model fits the data well (i.e. CFI = .973; TLI = .962 and RMSEA = .045). This means that respondents from different countries do not differ from one another in terms of the conceptual meaning attached to the six work environment factors. A further examination of the results presented in Table 5 reveals that the metric equivalence model also fits the data well (see goodness-of-fit statistics reported in Table 5). As the difference in CFI between the form equivalence model and the metric equivalence model falls below the critical difference of .01 (i.e. $.008 < .010$), the metric equivalence model may be preferred over the form equivalence model. This finding suggests that all respondents calibrate the intervals used on the measurement scale in similar ways, regardless of the country in which they are based. As scalar equivalence across countries was not established, (estimated) comparisons of factor mean scores across (all) countries may be misleading and, therefore, meaningless.

Table 5. Measurement equivalence of organizational survey across all countries

Comparison across all countries	Measurement equivalence model	χ^2 (*df*)	CFI	Difference in CFI (M1–M2 or M2–M3)	TLI	RMSEA
All 24 countries	M1: Form	6473.1 (1800)	.973	–	.962	.045
	M2: Metric	8053.4 (2007)	.965	.008	.956	.049
	M3: Scalar	15181.7 (2214)	.925	.040	.914	.068

Within cultural region measurement equivalence tests

Our hypotheses posited that measurement equivalence would be established across countries that are part of the same cultural region as identified by Schwartz. Consistent with the aforementioned results across all countries, Table 6 shows that form and metric equivalence of the survey instrument was supported within each cultural region as the model fit values did not exceed the critical values of Hu and Bentler (1999). This is not surprising as our previous analyses demonstrated that metric equivalence of the survey instrument was established across all 24 countries. Given that scalar equivalence was not established across all countries, it was especially interesting to investigate whether we found support for the scalar equivalence model in these within-cultural region analyses. Table 6 shows that scalar equivalence was better for the English-speaking countries with model fit statistics satisfying all criteria. Equivalence was somewhat worse in the other country groups, with fit statistics in most cases being a little below the criteria. Based on a statistical comparison between this model and the metric equivalence model (using the difference in CFI), scalar equivalence was accepted only in the English-speaking region (i.e. Australia, United Kingdom, United States and Canada). This means that the estimated factor means can be meaningfully compared across the English-speaking countries only. In the other cultural regions, a comparison of (estimated) factor means may be misleading, if not erroneous.

To examine the robustness of these results, we conducted three ancillary analyses. First, we re-ran the aforementioned analyses keeping sample size constant (see Table 2). Specifically, measurement equivalence of the survey instrument was tested using a random sample of 250 observations per country (or, alternatively, the original sample if the original sample did not exceed 250). Exactly the same results (available from the first

Table 6. Measurement equivalence of organizational survey within cultural regions

Comparisons within a cultural region	Measurement equivalence model	χ^2 (*df*)	CFI	Difference in CFI (M1–M2 or M2–M3)	TLI	RMSEA
West Europe (Seven countries, $N = 6273$)	M1: Form	1375.1 (525)	.971	–	.960	.045
	M2: Metric	1605.7 (579)	.966	.005	.956	.047
	M3: Scalar	2705.9 (633)	.930	.036	.919	.064
East Europe (Two countries, $N = 1424$)	M1: Form	365.4 (150)	.969	–	.956	.045
	M2: Metric	398.2 (159)	.965	.004	.954	.046
	M3: Scalar	528.4 (168)	.948	.017	.935	.055
English-speaking (Four countries, $N = 9149$)	M1: Form	2085.1 (300)	.972	–	.961	.051
	M2: Metric	2158.7 (327)	.971	.001	.963	.049
	M3: Scalar	2305.9 (354)	.969	.002	.964	.049
Latin countries (Five countries, $N = 11968$)	M1: Form	1836.0 (375)	.975	–	.966	.04
	M2: Metric	2192.5 (411)	.970	.005	.962	.043
	M3: Scalar	3884.6 (447)	.942	.028	.932	.057
Far East (Two countries, $N = 410$)	M1: Form	231.4 (150)	.960	–	.943	.051
	M2: Metric	244.6 (159)	.958	.002	.944	.051
	M3: Scalar	293.9 (168)	.938	.020	.922	.060
Islamic countries (Four countries, $N = 2108$)	M1: Form	579.9 (300)	.973	–	.963	.042
	M2: Metric	686.4 (327)	.966	.007	.956	.046
	M3: Scalar	1134.9 (354)	.925	.041	.911	.065

author) were obtained. Second, to evaluate whether differences in job level would be responsible for rejecting the scalar equivalence model (see, for instance, Fenlason & Suckow-Zimberg, 2006 and Roberts, Konczal, & Hoff Macan, 2004) in all but the English-speaking region, the within-regions measurement equivalence tests were repeated using random subsamples of all employees per country. These subsamples were identical in size and had an equal proportion of employees from each job level (i.e. low, intermediate and high) in all countries belonging to the cultural region under consideration. Results (available from the first author) showed that controlling for differences in job level did not alter our results. Third, we analysed whether the language used in the survey could be held (partly) responsible for the weak level of measurement equivalence (i.e. metric equivalence only) obtained in some cultural regions (e.g. West Europe). In these analyses, countries with (at least one) common language were compared to one another. As shown in Table 7, scalar equivalence was found across pairs of West European countries (i.e. Belgium vs. the Netherlands; Switzerland vs. Germany) in which a common language was used. Similar analyses were conducted using the data from the Spanish-speaking Latin/South American countries. Although no support was found for the highest measurement equivalence model (i.e. the scalar equivalence model) across these countries (i.e. based on all model fit criteria), the difference in model fit between the metric equivalence model and the scalar equivalence model (as measured by the difference in CFI) decreased substantially as soon as Brazil (i.e. the only country in which the spoken language is not Spanish but Portuguese) was excluded from the cultural cluster of Latin/South American countries. This implies that, even though scalar equivalence was not demonstrated, it becomes a more realistic assumption if only Spanish speaking countries are selected to represent the Latin/South American cultural region. Finally, measurement equivalence was also evaluated across two Islamic countries which were characterized by a common language (English) in the organizational survey (see Table 2). Despite the fact that our previous analyses within the Islamic cultural region only supported the assumption of metric equivalence, Table 7 indicates that there is evidence for scalar equivalence across these two Islamic countries.

Table 7. Comparison across countries within cultural region with at least one common language

Comparisons between countries	Measurement equivalence model		CFI	Difference in CFI (M1–M2 or M2–M3)	TLI	RMSEA
Belgium vs. The Netherlands	M1: Form	391.7 (150)	.968	–	.955	.045
	M2: Metric	410.6 (159)	.966	.002	.955	.045
	M3: Scalar	480.7 (168)	.958	.008	.947	.048
Switzerland vs. Germany	M1: Form	380.3 (150)	.969	–	.957	.047
	M2: Metric	401.8 (159)	.967	.002	.957	.047
	M3: Scalar	453.3 (168)	.962	.005	.952	.050
Spanish speaking countries (Argentina, Chile, Honduras, Mexico)	M1: Form	409.7 (220)	.954	–	.934	.065
	M2: Metric	462.1 (244)	.947	.007	.932	.066
	M3: Scalar	517.8 (268)	.939	.008	.929	.068
Malaysia vs. Pakistan	M1: Form	282.4 (150)	.970	–	.959	.050
	M2: Metric	298.8 (159)	.969	.001	.959	.050
	M3: Scalar	342.2 (168)	.961	.008	.951	.054

Between-cultural region measurement equivalence tests

To test our second set of hypotheses, measurement equivalence tests were also conducted between countries belonging to different cultural regions. As the data provided no evidence for scalar equivalence within cultural regions (except for the English-speaking region), aggregation of country-specific data to the regional level was not justifiable from a methodological point of view. Hence, analyses between cultural regions had to take into account all countries belonging to all cultural regions involved in the comparison. To be consistent in our data analytical approach we decided not to make an exception for the English-speaking cultural region. A summary of the results are presented in Table 8. Based on cultural distance as represented in Schwarz' coplot theory (see Schwartz, 1994, p. 58), combinations of cultural regions with high or low cultural distance were listed. For instance, the comparison of the West European region and the English-speaking region was considered to be low in cultural distance, whereas the comparison of the English-speaking region and the Far East region was considered to be high in cultural distance.

Table 8 shows that, between countries in cultural regions with a small cultural distance (according to Schwartz's coplot theory), our survey instrument consistently

Table 8. Comparison across cultural regions

Comparisons across cultural regions	Measurement equivalence model	χ^2 (*df*)	CFI	Difference in CFI (M1–M2 or M2–M3)	TLI	RMSEA
Large cultural distance expected						
East Europe vs. English-speaking (Six countries)	M1: Form	2450.4 (450)	.972	–	.960	.050
	M2: Metric	2694.6 (495)	.969	.003	.960	.050
	M3: Scalar	3189.6 (540)	.962	.007	.956	.053
East Europe vs. West Europe (Nine countries)	M1: Form	1740.6 (675)	.971	–	.959	.045
	M2: Metric	2124.7 (747)	.962	.009	.952	.049
	M3: Scalar	3634.3 (819)	.923	.039	.911	.066
English-speaking vs. Far East (Six countries)	M1: Form	2316.4 (450)	.972	–	.960	.051
	M2: Metric	2440.1 (495)	.970	.002	.962	.050
	M3: Scalar	2759.1 (540)	.966	.004	.961	.051
Far East vs. West Europe (Nine countries)	M1: Form	1606.6 (675)	.971	–	.959	.045
	M2: Metric	1878.3 (747)	.964	.007	.955	.048
	M3: Scalar	3064.0 (819)	.929	.035	.919	.064
English-speaking vs. Islamic Countries (Eight countries)	M1: Form	2665.0 (600)	.972	–	.961	.049
	M2: Metric	3011.2 (663)	.968	.004	.960	.050
	M3: Scalar	3859.8 (726)	.958	.010	.951	.055
West Europe vs. Islamic countries (Eleven countries)	M1: Form	1955.1 (825)	.972	–	.961	.044
	M2: Metric	2439.4 (915)	.962	.010	.952	.049
	M3: Scalar	4246.6 (1005)	.919	.043	.907	.068
Small cultural distance expected						
West Europe vs. English-speaking (Eleven countries)	M1: Form	3460.2 (825)	.972	–	.960	.049
	M2: Metric	3952.9 (915)	.967	.005	.959	.050
	M3: Scalar	6101.6 (1005)	.945	.022	.937	.061
Far East vs. Islamic countries (Six countries)	M1: Form	811.3 (450)	.971	–	.959	.044
	M2: Metric	951.2 (495)	.963	.008	.953	.047
	M3: Scalar	1529.4 (540)	.921	.042	.908	.066

showed metric equivalence. Metric equivalence was also found across countries in cultural regions with a large cultural distance (e.g. East Europe vs. West Europe; Far East vs. West Europe; West Europe vs. Islamic countries and West Europe vs. English-speaking region). In some comparisons between cultural regions with a large cultural distance, the highest form of equivalence, namely scalar equivalence was reported (e.g. East Europe vs. English-speaking region; Far East vs. English-speaking and English-speaking region vs. Islamic countries). As such, the data did not provide empirical evidence for the hypothesized (negative) relationship cultural distance (i.e. small or large) and the form of measurement equivalence between cultural regions.

Discussion

Main conclusions

This study aimed to examine the measurement equivalence of a survey instrument across 25 countries spread across all cultural regions in the world. Measurement equivalence was, however, assessed across only 24 countries as the data from South Africa did not meet basic psychometric conditions needed to perform measurement equivalence tests. Across these 24 countries, the survey instrument exhibited form and metric equivalence. These findings indicate that managers of the same organization use a similar frame-of-reference when completing items of an international survey. In addition, managers seem to calibrate the intervals of the rating scale in similar ways. These results corroborate previous multinational survey research of Ryan *et al.* (1999) and Liu *et al.* (2004) but extend them to countries (Egypt, Honduras, Pakistan, etc.) that have remained largely unexplored so far. However, there was no evidence for scalar equivalence across the 24 countries. Differences in response styles across countries (e.g. acquiescence bias) might explain why the mean-structure of the observed variables across countries was distorted. This possible explanation for the lack of evidence for scalar equivalence is in line with prior evidence from cross-cultural research (e.g. Byrne & Watkins, 2003; Cunningham *et al.*, 1977; Grimm & Church, 1999; Morris & Pavett, 1992; Riordan & Vandenberg, 1994; Ross & Mirowski, 1984; Van Herk *et al.*, 2004).

This study also tested whether Schwartz's (1994, 1999, 2004) theory of cultural values could serve as a viable theoretical framework for explaining why measurement equivalence can/cannot be established across countries and cultural regions. On the basis of Schwartz's theoretical framework, we formulated two set of hypotheses. One set of hypotheses dealt with the establishment of measurement equivalence within cultural regions. There was partial support for this hypothesis. In the English-speaking cultural region (Australia, Canada, United Kingdom and United States), we found empirical evidence for scalar equivalence of the survey instrument across countries. In all other cultural regions, fit statistics associated with the scalar equivalence model were a little below the required model fit criteria and the survey instrument exhibited only form and metric equivalence.

A second set of hypotheses was based on the notion of cultural distance. We hypothesized that measurement equivalence would be higher among countries that were part of cultural regions with a small cultural distance than among countries that were part of cultural regions with a large cultural distance. No support for this hypothesis was found. In contrast to our expectations, the highest form of measurement equivalence (i.e. scalar equivalence) was found across some pairs of cultural regions with a large cultural distance but never across regions with a small cultural distance.

A relevant question is how the establishment of scalar equivalence across pairs of cultural regions with a large cultural distance may be explained.

One possible explanation for this could hinge on the apparent preference of multinational organizations in Europe (cf. multinational organization of this study) to invest much more in aligning its workforce in countries from distant regions, rather than in countries from contiguous regions or in Europe. For example, a multinational corporation with headquarters in Europe might put much more efforts in promoting more Western values among its personnel in a plant in China than in a plant in East-Europe. As another explanation, some confounding effects due to language might have occurred so that the large cultural distance is reduced. A recent study by Harzing *et al.* (2005), for instance, has shown than cross-national differences may be underestimated if respondents answer in a common foreign language (English) instead of in their first language. In our study, Pakistanis and Malays responded primarily in English. As noted by an anonymous reviewer, their first languages are Urdu and Bahasa Melayu, respectively. As a broader explanation, it should be noted that some of the Schwartz's regions may be more heterogeneous than others. There is no strong empirical basis for the boundaries that Schwartz drew between the regions. In addition, variance within any of the regions is not considered to be non-negligible.

Apart from Schwartz's framework, this study also showed that use of a common language might contribute to or detract from the establishment of measurement equivalence. In fact, as the English-speaking cultural region was the only one with a common language (English), it seems plausible that this common language is responsible for the higher degree of measurement equivalence obtained in these countries. Further evidence for the interplay between language and culture was provided by our additional analyses in the regions wherein no evidence for scalar equivalence was found. Although such evidence was lacking in the West European cultural region, the survey data provided sufficient evidence for scalar equivalence of the survey instrument across countries which share the same language. Similar findings were obtained for the Islamic cultural region and to some extent also in the Latin/South American region.

Limitations

This study is not without limitations. First, we had no information on the response rate for each country. In addition, we made no corrections for demographic variations between the samples (with the exception of job level) because this study is based on a premise that is common in cross-cultural organizational research, namely that the country is the unit of analysis (Schaffer & Riordan, 2003). However, we acknowledge that workforces of adjacent nations might resemble one another more closely. Gender ratios, age structures, the types of organizational function and the types of business might all covary with membership of one or another Schwartz region. For instance, there are substantial differences between regions in terms of the percentage of managers working in production facilities (i.e. covering jobs in production, quality control and engineering). As shown in Table 2, this percentage is highest in East-Europe and the Latin/South American countries (i.e. the average ranking is 4.5 and 5.6, respectively; see Table 2). An intermediate position is taken by the English-speaking countries and Europe (i.e. the average ranking is 11.4 and 15.4, respectively). In the Islamic countries and the Far East, there aren't that many managers working in production facilities (i.e. average ranking is 20.0 and 21.5, respectively).

Second, some might argue that the organization specific survey under investigation in this study was not an established measure. Indeed, we used a company-specific instrument wherein each of the factors were measured with a limited number of items. We do not see the use of a customized instrument as a critical limitation. Consistent with Ryan *et al.* (1999), we believe that the essence of organizational surveying is that customized measures are often constructed that enable an organization to achieve its specific purposes. As mentioned above, the six constructs measured in this specific company's survey programme map well into the factors commonly included in organizational climate surveys (see James & Jones, 1974; Parker *et al.*, 2003). Nevertheless, it should be clear that future studies are needed to confirm our results in other organizations, in other countries, and with other measures.

Finally, when assessing measurement equivalence it was not possible to correct for possible response styles (e.g. acquiescence response style and extreme response style) across countries. Our data did not allow for such a correction. As indicated by Weijters (2006), adequate correction of response styles requires a specific set of items which are randomly selected from a wide variety of (multi-item) scales (see, for instance, the scales compiled by Robinson, Shaver, & Wrightsman, 1991). Such set of items was not included in our dataset as our data only included content-specific survey items. If response style indicators are derived on the basis of content-specific items, *content* and *response style* are confounded (e.g. Arce-Ferrer, 2006). As a consequence, any analytical procedure aimed at correcting for response styles (e.g. within-subject standardisation of item scores) would become invalid.

Practical implications

This study has also various implications for practice. Due to the internationalization of the economy and business environment, international surveying has become common practice. When developing an international survey, organizations typically decide to use an imposed-etic approach (Katigbak, Church, & Akamine, 1996; Triandis & Marin, 1983) for developing global surveys. This means that the original survey instrument developed in one culture is assumed to be universally applicable to all cultures. The rationale behind choosing an etic approach turns on the fact that it allows multinational organizations to quickly adapt HR practices in a global workforce (Ryan *et al.*, 1999). One of the challenges then concerns the importance of being able to make justified comparisons across the various countries on the basis of survey data. Such cross-country comparisons of survey data provide a global perspective on employee views and might help multinational organizations to differentiate between the various countries. In-turn, an understanding of cultural differences might encourage the modification of local work conditions so that they result in more favourable employee attitudes in specific countries. Basically, cross-country comparisons (in the context of benchmarking) are only meaningful from a substantive point of view if comparability of data is established across countries. If the survey scales used do not exhibit scalar equivalence across countries, comparisons between countries based on their (estimated) construct mean scores and the resulting interventions might be inaccurate.

Our findings suggest that, within cultural regions, the use of a common language in an organizational survey may reduce the bias present in cross-country comparisons. As mentioned before, a reduction of (cross-country) bias as present in the data from organizational surveys is critical as it may enable researchers to compare (estimated) construct means across factors may be responsible for the failure to establish scalar equivalence of the survey instrument.

Directions for future research

With respect to future research, we believe the following two avenues deserve further attention. First, it seems useful to compare measurement equivalence tests *across* as well as *within* countries. As noted above, prior research on the measurement equivalence of international surveys has generally assumed within-country homogeneity (Ryan *et al.*, 1999). That is, the survey data from various business units within a given country are aggregated at the country level. However, it is important to obtain empirical evidence to support this aggregation of data. It might well be that the within-country variation across business units is at least as large as the across-country variation.

Second, future studies should focus on *across-time* comparisons. Surveys and other instruments are not only used to compare countries on factors of interest. In addition, there is increasing interest in international cross-country comparisons across time. To the best of our knowledge, no research has addressed this issue. However, latent growth modelling (see, for instance, Muthén, 2004) could be fruitfully used to ensure that there is a sound psychometric basis to conduct these across-time comparisons.

Acknowledgement

The authors are indebted to Paul E. Spector and Frederik Anseel for their useful comments on an earlier version of this paper.

References

Arce-Ferrer, A. J. (2006). An investigation into the factors influencing extreme response style. *Educational and Psychological Measurement*, *66*, 374–392.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*, 238–246.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structure. *Psychological Bulletin*, *88*, 588–606.

Brannick, M. T. (1995). Critical comments on applying covariance structure modeling. *Journal of Organizational Behavior*, *16*, 201–213.

Byrne, B. M., & Watkins, D. (2003). The issue of measurement equivalence revisited. *Journal of Cross-Cultural Psychology*, *34*, 155–175.

Chan, D. (2000). Detection of differential item functioning on the Kirton adaption-innovation inventory using multi-group mean and covariance structure analyses. *Multivariate Behavioral Research*, *35*, 169–1999.

Cheung, G. W., & Rensvold, R. B. (2000). Assessing extreme and acquiescence response sets in cross-cultural research using structural equations modeling. *Journal of Cross-Cultural Psychology*, *31*, 187–212.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement equivalence. *Structural Equation Modeling: An Interdisciplinary Journal*, *9*(2), 233–255.

Chun, K. T., Campbell, J. B., & Yoo, J. H. (1974). Extreme response style in cross-cultural research: A reminder. *Journal of Cross-Cultural Psychology*, *5*, 465–480.

Church, A. H., & Waclawski, J. (2001). *Designing and using organizational surveys: A seven-step process*. San Francisco, CA: Jossey-Bass.

Clarke, I. (2000). Extreme response style in cross-cultural research: An empirical investigation. *Journal of Social Behavior and Personality*, *15*, 137–152.

Cole, M. S., Bedeian, A. G., & Field, H. S. (2006). The measurement equivalence of web-based and paper-and-pencil measures of transformational leadership. *Organizational Research Methods*, *9*, 339–368.

Cunningham, W., Cunningham, W. C. M., & Green, R. T. (1977). The ipsative process to reduce response set bias. *Public Opinion Quarterly*, *41*, 379–394.

Drasgow, F. (1984). Scrutinizing psychological tests: Measurement equivalence and equivalent relations with external variables are central issues. *Psychological Bulletin*, *95*, 134–135.

Drasgow, F. (1987). Study of measurement bias of two standardized psychological tests. *Journal of Applied Psychology*, *72*, 19–29.

Fenlason, K. J., & Suckow-Zimberg, K. (2006). Online surveys. Critical issues in using the web to conduct surveys. In A. I. Kraut (Ed.), *Getting action from organizational surveys* (pp. 183–212). San Franciso, CA: Jossey-Bass.

Gerbing, D. W., & Hamilton, J. G. (1996). Viability of exploratory factor analysis as a precursor to confirmatory factor analysis. *Structural Equation Modeling. An Interdisciplinary Journal*, *3*, 62–72.

Grimm, S. D., & Church, A. T. (1999). A cross-cultural study of response biases in personality measures. *Journal of Research in Personality*, *33*, 415–441.

Harzing, A.-W., Fischlmayr, I., Freitas, M. E., Lazarova, M., Liberman Yaconi, L., Zhu, Y., *et al.* (2005). Does the use of English-language questionnaires in cross-national research obscure national differences? *International Journal of Cross Cultural Management*, *5*, 213–224.

Hofstede, G. (1980). *Culture's consequences: International differences in work-related values*. Beverly Hills: Sage.

Hofstede, G. (1991). *Cultures and organizations: Software of the mind*. London: McGraw Hill.

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

Hulin, C. L., & Mayer, L. J. (1986). Psychometric equivalence of a translation of the job description index into Hebrew. *Journal of Applied Psychology*, *71*, 83–94.

Hurley, A. E., Scandura, T. A., Schriesheim, C. A., Brannick, M. T., Seers, A., Vandenberg, R. J., *et al.* (1997). Exploratory and confirmatory factor analysis: Guidelines, issues, and alternatives. *Journal of Organizational Behavior*, *18*, 667–683.

James, L. R., & Jones, A. P. (1974). Organizational climate: A review of theory and research. *Psychological Bulletin*, *81*, 1096–1112.

James, L. R., & McIntyre, M. D. (1996). Perceptions of organizational climate. In K. Murphy (Ed.), *Individual differences and behavior in organizations* (pp. 416–450). San Francisco: Jossey-Bass.

James, L. R., & Sells, S. B. (1981). Psychological climate: Theoretical perspectives and empirical research. In D. Magnusson (Ed.), *Toward a psychology of situations: An interactional perspective* (pp. 275–295). Hillsdale, NJ: Erlbaum.

Johnson, S. R. (1996). The multinational opinion survey. In A. I. Kraut (Ed.), *Organizational surveys: Tools for assessment and change* (pp. 310–329). San Francisco: Jossey-Bass.

Johnson, T., Kulesa, P., Cho, Young, Ik., & Shavitt, S. (2005). The relation between culture and response styles: Evidence from 19 countries. *Journal of Cross-Cultural Psychology*, *36*, 264–277.

Katigbak, M. S., Church, A. T., & Akamine, T. X. (1996). Cross-cultural generalizability of personality dimensions: Relating indigenous and imported dimensions in two cultures. *Journal of Personality and Social Psychology*, *70*, 99–114.

Kelloway, E. K. (1995). Structural equation modeling in perspective. *Journal of Organizational Behavior*, *16*, 215–224.

Kluckhohn, F. R., & Strodtbeck, F. L. (1961). *Variations in value orientations*. Evanston, IL: Row, Peterson.

Kogut, B., & Singh, H. (1988). The effect of national culture on the choice of entry mode. *Journal of International Business Studies*, *19*, 411–432.

Kraut, A. I. (1996). *Organizational surveys: Tools for assessment and change*. San Francisco, CA: Jossey-Bass.Little.

Kraut, A. I. (2006). Moving the needle. Getting action after a survey. In A. I. Kraut (Ed.), *Getting action from organizational surveys* (pp. 1–30). San Franciso, CA: Jossey-Bass.

Lee, C., & Green, R. T. (1991). Cross-cultural examination of the Fishbein behavioral intentions model. *Journal of International Business Studies*, *22*, 289–305.

Little, T. D. (1997). Mean and covariance (MACS) analysis of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioural Research*, *32*, 53–76.

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.

Manev, I. M., & Stevenson, W. B. (2001). Nationality, cultural distance, and expatriate status: Effects on the managerial network in a multinational enterprise. *Journal of International Business Studies*, *32*, 285–303.

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

Morris, T., & Pavett, C. M. (1992). Management style and productivity in two cultures. *Journal of International Business Studies*, *23*, 169–179.

Muthén, B. (2004). Latent variable analysis: Growth mixture modeling and related techniques for longitudinal data. In D. Kaplan (Ed.), *Handbook of quantitative methodology for the social sciences*. Newbury Park, CA: Sage.

Muthén, L. K., & Muthén, B. O. (1999). *Mplus, the comprehensive modeling program for applied researchers (User's guide – 2nd printing)*. Los Angeles, CA: Muthén and Muthén.

Muthén, L. K., & Muthén, B. O. (2003). *Mplus version 2.13 Addendum to the Mplus user's guide*. Los Angeles, CA: Muthén and Muthén.

Netmeyer, R. G., Bearden, W. O., & Sharma, S. (2003). *Scaling procedures. Issues and Applications*. Thousand Oaks, CA: Sage.

Parker, C. P., Baltes, B. B., Young, S. A., Huff, J. W., Altmann, R. A., Lacost, H. A., *et al.* (2003). Relationships between psychological climate perceptions and work outcomes: A meta-analytic review. *Journal of Organizational Behavior*, *24*, 389–416.

Riordan, C. R., & Vandenberg, R. J. (1994). A central question in cross-cultural research: Do employees of different cultures interpret work-related measures in an equivalent manner? *Journal of Management*, *20*, 643–671.

Roberts, L. L., Konczal, L. J., & Hoff Macan, T. (2004). Effects of data collection method on organizational climate survey results. *Applied H.R.M. Research*, *9*, 13–26.

Robinson, J. P., Shaver, P. R., & Wrightsman, L. S. (1991). *Measures of personality and social psychological attitudes* (Vol. 1). London, UK: Academic Press.

Ross, C. E., & Mirowski, J. (1984). Socially desirable response and acquiescence in a cross-cultural survey of mental health. *Journal of Health and Social Behavior*, *25*, 189–197.

Rucci, A. J., Kirn, S. P., & Quinn, R. T. (1998). The employer-customer-profit chain and sears. *Harvard Business Review*, 82–97.

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

Saari, L. M., & Judge, T. A. (2004). Employee attitudes and job satisfaction. *Human Resource Management*, *43*, 395–407.

Schaffer, B. S., & Riordan, C. M. (2003). A review of cross-cultural methodologies for organizational research: A best-practices approach. *Organizational Research Methods*, *6*(2), 169–215.

Schwartz, S. H. (1994). *Beyond individualism/collectivism: New cultural dimensions of values. Individualism and collectivism: Theory, method, and applications*. Thousand Oaks, CA: Sage.

Schwartz, S. H. (1999). A theory of cultural values and some implications for work. *Applied Psychology: An International Review*, *48*, 23–48.

Schwartz, S. H. (2004). Mapping and interpreting cultural differences around the world. In H. Vinken, J. Soeters, & P. Ester (Eds.), *Comparing cultures: Dimensions of culture in a comparative perspective* (pp. 43–73). Leiden, NL: Brill Academic Publishers.

Schwartz, S. H., & Bardi, A. (1997). Influences of adaptation to communist rule on value priorities in Eastern Europe. *Political Psychology*, *18*, 385–410.

Schwartz, S. H., & Ros, M. (1995). Values in the West: A theoretical and empirical challenge to the individualism-collectivism dimension. *World Psychology*, *1*, 91–122.

Schwarzer, R., Born, A., Iwawaki, S., Lee, Y. -M., Saito, E., & Yue, X. (1997). The assessment of optimistic self-beliefs: Comparison of the Chinese, Indonesian, Japanese and Korean versions of the General self-efficacy scale. *Psychologia: An International Journal of Psychology in the Orient*, *40*, 1–13.

Shenkar, O. (2001). Cultural distance revisited: Towards a more rigorous conceptualization and measurement of cultural differences. *Journal of International Business Studies*, *32*, 519–535.

Spector, P. E., Cooper, C. L., Sanchez, J. I., O'Driscoll, M., Sparks, K., Bernin, P., *et al.* (2002). A 24 nation/territory study of work locus of control in relation to well-being at work: How generalizable are western findings? *Academy of Management Journal*, *45*, 453–466.

Stanton, J. M. (1998). An empirical assessment of data collection using the internet. *Personnel Psychology*, *51*, 709–725.

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research*, *25*, 173–180.

Stening, B. W., & Everett, J. E. (1984). Response styles in a cross-cultural managerial study. *Journal of Social Psychology*, *122*, 151–156.

Triandis, H. C., & Marin, G. (1983). Etic plus emic versus pseudoetic. A test of a basic assumption of contemporary cross-cultural psychology. *Journal of Cross-Cultural Psychology*, *14*, 489–500.

Trompenaars, A., & Hampden-Turner, C. (1998). *Riding the waves of culture: Understanding cultural diversity in global business*. New York, NY: McGraw Hill.

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

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

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

Van Herk, H., Poortinga, Y. H., & Verhallen, T. M. M. (2004). Response styles in rating scales: Evidence of method bias in data from six EU countries. *Journal of Cross-Cultural Psychology*, *35*, 346–360.

Weijters, B. (2006). *Response styles in consumer research*. Unpublished doctoral dissertation, University of Ghent, Belgium.

Zax, M., & Takahashi, S. (1967). Cultural influences on response style: Comparisons of Japanese and American college students. *Journal of Social Psychology*, *71*, 3–10.

Received 17 January 2006; revised version received 18 December 2006

Appendix: Definitions and measures of the 6 factors

Factor 1: Team commitment

Definition: The extent to which employees of a department are working together towards a common objective by effectively exchanging information and by being dedicated to get the job done.

Items:

- In my department, people provide each other with useful feedback.[a]
- In my department, people usually do what they say they will.
- In my department, people do not accept mediocrity in their work.

Factor 2: Supervisor support

Definition: The extent to which employees perceive that supervisors help them in accomplishing their goals by providing feedback and information.

Items:

- My immediate boss gives me regular feedback on my performance.[a]
- My immediate boss communicates clearly.
- I feel my immediate boss coaches me when I need it.

Factor 3: Goal clarity

Definition: The extent to which employees know what is expected of them and how these role expectations translate into the goals and strategy of the organization.

Items:

- I have a clear understanding of the goals and objectives of my department.[a]
- I have a clear understanding of the goals and objectives of my organization.
- I have a clear understanding of the goals and objectives of the multinational as a whole.

Factor 4: Decision-making

Definition: The extent to which employees have confidence in the decisions made by direct supervisors and higher level managers.

Items:

- I have confidence in the decisions made by managers of my organization.[a]
- I have confidence in the decisions made by managers of my business group/region.

Factor 5: Organizational adaptability

Definition: The extent to which employees perceive that the organization and its members quickly adapt their practices, processes, and routines to sudden changes in the environment and market.

Items:

- In your judgment, how does this organization compare with its competitors on responding rapidly to changes in the market?[a]
- How good are managers in your organization doing in developing simple and fast processes from supplier through to consumer?

Factor 6: Environmental and societal responsibility

Definition: The extent to which employees perceive the organization to adopt business practices that embody environmental protection and responsibility to the society.

Items:

- I believe that my organization is environmentally responsible.[a]
- I believe that my organization is a socially responsible member of the community.

Note. [a] This item was arbitrarily chosen as reference indicator in the analyses.