4-2009

Measurement equivalence of paper-and-pencil and internet organisational surveys: A large scale examination in 16 countries

Alain DE BEUCKELAER
Radboud University Nijmegen

Filip LIEVENS
Singapore Management University, filiplievens@smu.edu.sg

Published in Applied Psychology, 2009 April,
Volume 58, Issue 2, Pages 336-361
http://doi.org/10.1111/j.1464-0597.2008.00350.x

Measurement Equivalence of Paper-and-Pencil and Internet Organisational Surveys: A Large Scale Examination in 16 Countries

Alain De Beuckelaer*
Radboud University Nijmegen, The Netherlands

Filip Lievens
Ghent University, Belgium

In multinational surveys, mixed-mode administration modes (e.g. combining Internet and paper-and-pencil administration) are increasingly used. To date, no studies have investigated whether measurement equivalence exists between Internet data collection and data collection using the conventional paper-and-pencil method in organisational surveys which include a large number of countries. This paper examined the measurement equivalence of a truly global organisational survey across Internet and paper-and-pencil survey administrations. Data from an organisational survey in 16 countries (N = 52,461) across the globe were used to assess the measurement equivalence of an organisational climate measure within each country in which the survey was administered. The empirical data provided strong indications which support the measurement equivalence of the multi-item survey instrument across Internet and paper-and-pencil surveys in virtually all countries in which the survey was conducted. These findings suggest that merging data obtained through Internet and paper-and-pencil data administration in a particular country is legitimate as no evidence was found for differential effects across both modes of data collection.

Dans les enquêtes internationales, les modalités d'administration mixtes, c'est-à-dire combinant les solution Internet et papier-crayon, sont de plus en plus utilisées. Jusqu'à présent, on ignore si le recueil de données par Internet est méthodologiquement équivalent au recueil traditionnel sur papier dans les enquêtes organisationnelles qui couvrent un grand nombre de pays. Dans cet article, on analyse l'analogie des mesures d'une enquête organisationnelle véritablement universelle utilisant à la fois une administration Internet et papier-crayon. Des données provenant de seize pays (N = 52,461) répartis

* Address for correspondence: Alain De Beuckelaer, Radboud University Nijmegen, Thomas van Aquinostraat 1, P.O. Box 9108, 6500 HK Nijmegen, The Netherlands. Email: A.DeBeuckelaer@fm.ru.nl

The authors would like to thank Martine van Sehm, Csilla Horvath, Olivier Furrer and three anonymous reviewers for their useful comments on earlier versions of this paper.

sur l'ensemble de la planète ont été exploitées pour évaluer l'équivalence méthodologique d'une mesure du climat organisationnel dans chacun de ces pays. Les résultats empiriques sont en faveur de l'équivalence méthodologique de l'ensemble des items pour les recueils Internet et papier-crayon dans pratiquement tous les pays retenus pour l'enquête. Ces résultats montrent qu'il est légitime de traiter ensemble les données obtenues par Internet et par papier-crayon dans un même pays puisque rien ne nous permet d'affirmer l'existence d'un effet différentiel dû aux deux modes d'administration.

INTRODUCTION

The globalisation of the marketplace is arguably the most important challenge companies face today. The globalisation process affects all aspects of policy-making in multinational companies, including Human Resource (HR) management. Global organisational surveys (also known as employee attitude surveys or employee opinion surveys) constitute a specific instrument which allows HR professionals to evaluate whether the global human resource management policy is effective. Given their global reach, organisational surveys are typically distributed to a wide variety of employees working in different countries all over the world.

Nowadays, organisations are increasingly making use of a "mixed-mode" administration approach. That is, organisational surveys typically combine survey administration using the Internet (i.e. web-based surveys) and its conventional equivalent, namely survey administration by means of the paper-and-pencil (PP) method. It is expected that this particular type of mixed-mode survey will account for a substantial proportion of organisational surveys in the foreseeable future (Fenlason & Suckow-Zimberg, 2006). In mixed-mode surveying, the choice of a specific data collection method is typically an individual (e.g. respondents have the discretion to choose a specific data collection method) and/or organisational decision (e.g. in some divisions and departments, Internet access is restricted by the organisation to specific groups of employees).

A key methodological assumption underlying such mixed-mode organisational surveys is that data from paper-and-pencil and Internet surveys can be meaningfully combined and compared. That is, to enable a meaningful substantive interpretation of the overall results, measurement equivalence across the two modes of data collection (e.g. Fenlason & Suckow-Zimberg, 2006, pp. 185–187) should be established. The establishment of measurement equivalence implies the absence of measurement effects (i.e. biases) of collecting survey data through the Internet as compared to using the PP method.

The purpose of this paper is to examine whether survey questions used to measure work-related constructs exhibit measurement equivalence across

PP and Internet administrations of a global organisational survey. Although prior studies have examined the measurement equivalence across PP and Internet administrations within a given country (Stanton, 1998; Buchanan & Smith, 1999), this study is the first one to examine measurement equivalence of a mixed-mode administration format in a large number of countries.

STUDY BACKGROUND

Paper-and-Pencil, Internet, and Mixed-Mode Organisational Surveys

In recent years, an increasing number of organisations have started to use the Internet as a medium for collecting data (Thompson, Surface, Martin, & Sanders, 2003). Generally, it has been argued that Internet surveying has several advantages over the more conventional PP method (Yun & Trumbo, 2000). As compared to PP surveys, Internet surveys are less costly (Dillman, 2000; Kraut & Saari, 1999; Schaeffer & Dillman, 1998; Sproull, 1986; Yun & Trumbo, 2000), lead to faster survey responses (Schaeffer & Dillman, 1998; Sproull, 1986), allow for greater flexibility in survey design (Dillman, 2000), and offer a wider variety of response formats (Simsek & Veiga, 2001). In addition, Internet surveys have a wider geographical reach (Epstein, Klinkenberg, Wiley, & McKinley, 2001), do not suffer from human (coding) errors (Cook, Heath, Thompson, & Thompson, 2001; Roberts, Konczak, & Macan, 2004), are free of experimenter bias (Birnbaum, 2001; Reips, 2000), are less sensitive to order of question effects due to the ease of randomising questions (see Bowling, 2005), and do not have that many missing values (i.e. they are more "complete"; Stanton, 1998).[1] As far as organisational surveys are concerned, the most important reasons for using Internet surveying (in addition to paper-and-pencil surveys) may be: (1) increased efficiency of the data collection process (e.g. faster data collection through the use of available PC technology), (2) the elimination of human coding errors, and (3) a cost-reduction (at least, if it concerns a truly large survey in many regions of the world).

However, Internet surveying has some drawbacks as well. Disadvantages of Internet surveying may include higher non-response rates (Schaeffer & Dillman, 1998; Sproull, 1986), a higher probability of getting dishonest answers (Booth-Kewley, Edwards, Rosenfeld, 1992; Lautenschlager & Flaherty, 1990), potential technological problems (Kraut & Saari, 1999),

[1] Some of the advantages of Internet surveying mentioned in the text may not always apply. Initial start-up costs, for instance, may reduce or even completely eliminate anticipated cost-savings. Similarly, technical problems may cause additional delays in the completion of the data collection process (Mann & Stewart, 2000; van Selm & Jankowski, 2006).

decreased item reliability due to somewhat higher measurement errors (Stanton, 1998), and the possibility of multiple submissions (Reips, 2000). In addition, Internet survey methods are somewhat restricted in that they may not provide sufficient coverage of all occupational groups represented in the organisation being studied (Bowling, 2005). As indicated by Stanton and Rogelberg (2001), individuals belonging to different occupational groups may have varying access to computers, e-mail, the Internet (or the Intranet) during working hours. Employees from different occupational groups may also differ from one another in terms of their computer literacy. Furthermore, regional differences may also exist as the use of Internet surveys may not be feasible in some specific regions in the world (e.g. in some of the developing countries) in which the organisation has business activities.

Given the limitations of Internet surveying just mentioned, it is often necessary to combine Internet surveys with more conventional modes of data collection such as the PP method. As such, the response rate of the total sample might improve (Yun & Trumbo, 2000). As noted above, if multinational organisations want to compare survey data from such mixed-mode (data collection) procedures, it is crucial that there are no measurement effects of collecting survey data through Internet as compared to using the PP method. In fact, establishing measurement equivalence across survey administration modes is necessary to enable meaningful interpretation of the substantive results in all countries participating in the research (see, for instance, Vandenberg, 2002; Vandenberg & Lance, 2000).

Research on Measurement Equivalence of Survey Measures across Internet and Paper-and-Pencil Administrations

In the past, various studies examined whether survey questions used to measure work-related constructs exhibit measurement equivalence across Internet and PP surveys. As shown in Table 1, prior studies used a wide variety of surveys including (but not limited to) organisational surveys. Note that Table 1 presents only those studies that used "advanced" analytical methods (i.e. confirmatory factor analysis and item response models) for testing for measurement equivalence across data collection modes.

One group of studies examined measurement equivalence of general psychological measures (thus not in an organisational context). Buchanan and Smith (1999) and Ferrando and Lorenzo-Seva (2005) relied on samples of students in the US and Spain, respectively. In the former study, the survey instrument comprised items measuring self-monitoring, whereas the latter survey instrument contained items of the Eysenck Personality Questionnaire. Both studies demonstrated measurement equivalence across different data

TABLE 1
Overview of Empirical Studies Investigating the Equivalence of Construct Measurement across Multiple Modes of Data Collection

Study	*Subject and countries involved*	*Constructs involved*	*Modes of data collection involved*	*Method to assess ME*	*Most relevant conclusions*
Stanton (1998)	Professional employees in the US (N = 50 Internet; N = 181 PP)	Organisational justice and consistency of supervision	Internet and PP	Multigroup CFA (without MS)	No serious violations of ME reported (but small differences in indicator reliabilities!)
Buchanan & Smith (1999b)	Students in the US (N = 963 Internet; N = 224 PP)	Self-monitoring	Internet and PP	Multigroup CFA (without MS)	No serious violations of ME reported
Vispoel, Boo, & Bleiler (2001)	University students in the US (N = 224 Internet; N = 224 PP)	Self-esteem	Computerised test administration and PP	Multigroup CFA (without MS)	No serious violations of ME reported
Ployhart, Weekley, Holtz, & Kemp (2003)	Job applicants in the US (N = 2,356 Internet; N = 2,544 PP and current employees (N = 425 PP)	Personality data, biodata, and situational judgment tests	Internet	Multigroup CFA (without MS)	Between-mode differences in terms of mean (Internet: lower means) and variance structures (Internet: larger variances), internal consistency measures (Internet: higher), and item correlations (Internet: stronger correlations)
Ferrando & Lorenzo-Seva (2005)	University students in Spain (N = 201 Internet; N = 201 PP)	Personality	Internet and PP	Multigroup CFA (without MS) and an IRT (DIF)-based approach	No serious violations of ME reported
Cole, Bedeian, & Feild (2006)	Employees from multiple countries	A measure of transformational leadership	Internet and PP	Multigroup CFA (with MS)	No violations of ME reported

Note: "PP" refers to "paper-and-pencil"; CFA refers to "confirmatory factor analysis"; MS refers to "means structures" (i.e. mean values for the indicator variables).

collection modes. Similar conclusions were drawn by Vispoel, Boo and Bleiler (2001). Conversely, Ployhart, Weekley, Holtz, and Kemp (2003) reported some differences across Internet surveys and PP surveys using psychological measures and other measures (e.g. biodata) in terms of mean structures, variance structures, internal consistency measures, and item correlations. Another study by Stanton (1998) examined measurement equivalence in an organisational survey context. US employees completed a survey related to fairness and organisational justice perceptions. Data from employees who responded via the Internet ($N = 50$) were compared to data from a comparable sample of employees ($N = 181$) who used the PP method. Apart from small differences in item reliabilities, confirmatory factor analyses revealed no measurement effects across data collection modes.

Although the results of these prior studies are insightful, they share several limitations. First, most studies relied on survey data from only one country (i.e. either the US or Spain). Accordingly, these studies are relevant only for surveys conducted in a national context and do not generalise to international surveys. The study by Cole, Bedeian, and Field (2006) is an exception as it had an international scope. Unfortunately, the study offered only a rough tentative test on possible violations of measurement equivalence of the survey instrument across Internet and PP surveys. Cole et al. (2006) had to rely on two aggregate samples (i.e. an Internet sample [$N = 4{,}244$] and a PP sample [$N = 665$]), each of which included survey responses from a large number of countries (50) in unequal proportions. As the Internet and PP samples were not homogeneous, their approach to measurement equivalence testing may have led to a biased comparison between the Internet and the PP sample. A related issue concerns the size of the sample used in Cole et al. (2006). As the PP sample by Cole et al. (2006) had only 13 observations per country on average (i.e. $N = 665/50$), it did not allow for making valid measurement equivalence tests between Internet and PP surveys at the country level. This is a serious limitation as most organisational surveys are conducted in a large number of countries, and results are typically reported at country level. Hence, country-specific assessments of systematic differences between Internet and PP surveys would be a welcome addition to the literature on organisational surveys. To the best of our knowledge, this study will be the first to make a large number of country-specific comparisons across Internet and PP surveys using comparable Internet and PP samples. The examination of mixed-mode differences within a large number of countries is critical as one should account for the possibility of cross-country differences in the establishment of measurement equivalence across PP and Internet administrations. In particular, a recent study in the marketing domain (Weijters, 2006) demonstrated that in Internet surveys respondents provide a "no answer" less frequently to an item (regardless of

item content). In addition, in Internet surveys respondents tend to use the extreme points of a scale less frequently. Cross-cultural research has shown that response tendencies (e.g. choosing the extreme rating points) differ from culture to culture (Johnson, Kulesa, Ik Cho, & Shavitt, 2005; Van Herk, Poortinga, & Verhallen, 2004). Therefore, it is crucial to assess measurement equivalence across both modes of data collection within a wide variety of countries.

Most previous studies are also limited because they did not use organisational members from different occupational groups (see Table 1). In most studies, students, job applicants, or employees were used to test for measurement equivalence between PP and Internet surveys. It is very likely that higher-level organisational members may show a stronger tendency to respond via the Internet than lower-level organisational members. The reason for this may be that the former group has more frequent access to the Internet than the latter group. Hence, an adequate assessment of measurement equivalence of an organisational measure should take into account organisational members' job level.

A third limitation relates to the use of the analytical technique for testing measurement equivalence. In all prior studies (with the exception of Cole et al., 2006), multigroup confirmatory factor analysis (CFA) was used. In CFA, only covariances between observed variables (i.e. items) are modeled. Mean values of the observed variables (items) are not modeled. As such, multigroup CFA may detect differences in factor loadings of survey items across modes of data collection (i.e. the groups under study). Yet, it fails to detect differences in item intercepts. The equality of factor loadings *and* item intercepts (also known as *scalar equivalence*) is required whenever (estimated) construct means are to be compared across groups or countries (see Chan, 2000; Little, 1997; Meredith, 1993). To test for scalar equivalence, using a multigroup Mean and Covariance Structure (MACS) analysis is generally recommended (Sörbom, 1974, 1978). A MACS analysis differs from a multigroup CFA in that item intercepts are modeled in addition to the covariances between the items.

Purpose and Contributions of Present Study

The purpose of this study is to examine the measurement equivalence of survey data across modes of data collection (PP and Internet) in mixed-mode surveys conducted in a large number of countries. This study contributes to the existing literature on survey research in three key ways. First, we examine whether the data from a truly global organisational survey show measurement equivalence across modes of data collection using data from *multiple countries* (16) based on disaggregate observations. As argued before, measurement invariance tests are conducted on a country-by-country

basis. Second, we examine the measurement equivalence across data collection modes while controlling for differences in employees' *job level*. Third, we use *Mean and Covariance Structure Analysis* to formally test the assumption of *scalar* equivalence as part of the series of measurement equivalence tests conducted.

METHOD

Sample and Procedure

Data from 52,461 respondents in 16 countries were collected in 2004 within a multinational company in the fast-moving consumer goods sector. The multinational has business operations in more than 70 countries across the world. The actual survey contains data on more than 16 countries but countries were only selected for inclusion in this study if the sample size of Internet and PP responses exceeded $N = 90$. As shown later in this paper, this minimum sample size resulted in a subject–variable ratio of about 7.0.

In principle, all employees (with different job levels) were surveyed. In the United States, however, some lower-level employees were not surveyed (due to union restrictions). 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 organisational survey was (1) to enhance employees' involvement and motivation and (2) to provide baseline data for organisational change efforts. Data collection through the Internet started after all employees having regular access to the Internet received a mass mail. They responded to the survey at their individual work site. Reminders were sent to people who did not respond within a period of two weeks after data collection had started, and just before closing the survey administration phase. Human Resource professionals made hardcopies of the survey available to those employees who had no access to the Internet at their work site or to employees who explicitly asked to receive a hardcopy. In this organisational survey, employees providing their responses on paper were typically employees with the lowest job level working: (1) in sales (77% versus 23%), (2) in manufacturing, maintenance, quality and engineering (90% versus 10%), and (3) in warehouses and coldstores (88% versus 12%). Most of them did not have access to the Internet during working hours. Next, most top-level managers (e.g. CEOs in countries, business group directors, etc.) also provided their answers on paper (63%). Responding through the Internet was popular among all other types of employees. The average percentage of Internet and paper-and-pencil responses across countries was 35.6 per cent and 64.4 per cent, respectively. The lowest percentage of Internet responses was found in Vietnam (i.e. 9.1%), whereas the highest percentage of Internet responses was found in the United States (i.e. 60.2%).

Across countries the average percentage of employees having the lowest job level was 81.7 per cent (SD = 7.9). These employees do not have any staff reporting to them. The average percentage of employees occupying an intermediate job level (i.e. team leaders or head of a department) across countries was 12.3 per cent (SD = 5.8). The average percentage of employees with a high job level (i.e. all positions above head of the department) was 6.0 per cent (SD = 2.7).

In total, data collection took about four months. The overall response rate across countries was 86.0 per cent, which is above the average survey response rate given by Church and Waclawski (2001) and Kraut (1996). Unfortunately, no response rates per country were made available to the authors.

Development of Organisational Survey

The organisational survey under investigation in this study was constructed analogously to corporate-sponsored global surveys (see Johnson, 1996). This means that the HR staff at the corporate headquarters led the development and administration of the survey. In addition, a common methodology and framework to survey all employees was followed, while allowing for country customisation. In a first step, a global survey team was composed. 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 original English items. Some of the items had been used before by the survey consultants. Others were added by the global survey team. The closed-ended items of the survey used a 5-point Likert-type response format. The response format ranged from "*strongly disagree*" (1) to "*strongly agree*" (5). In the following step, regional survey leaders were made responsible for the translations in their region. They supervised and monitored the different translations which were checked by local survey coordinators using the English survey as the basis for comparison. Next, professional interpreters back-translated the surveys into English. As an additional check, English-speaking masters students 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 included 89 items and consisted of several parts. One part asked for organisational members' reports and perceptions on dimensions which were of key importance to the HR policy of the multinational company. Specifically, the following five organisational climate factors were considered to be of key importance: team commitment, supervisor support,

goal clarity, decision-making, and environmental and societal responsibility. The Appendix presents the definitions of these organisational climate factors. A second part of the survey included country-specific questions, whereas the last part dealt with background information (e.g. age, gender, tenure).

In this study, we focused on a specific set of items (13) of this global survey (see Appendix). Consistent with Ryan, Chan, Ployhart, and Slade (1999), we retained only items that were clearly linked[2] to the five work climate factors. Hence, items dealing with either country-specific or organisation-specific initiatives (instead of attitudes or opinions) were not retained.

Although the five work climate factors measured in this organisational survey were specific to this company, two comments are in order. First, it should be noted that they generalise well to factors typically included in organisational climate surveys (James & Jones, 1974; Kopelman, Brief, & Guzzo, 1990; Ostroff, 1993). For example, a well-known comprehensive model of organisational climate (James & Jones, 1974; James & Sells, 1981; James & McIntyre, 1996; Parker, Baltes, Young, Huff, Altmann, Lacost, & Roberts, 2003) makes a distinction between five primary domains of work climate: job characteristics, role characteristics, leadership characteristics, work group/social characteristics, and organisational characteristics. The organisational survey of this study measured four of these five core organisational 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 organisational characteristics (i.e. environmental and societal responsibility). Only job characteristics were not measured in this study's organisational survey.

---

[2] We began by screening the 89 survey items in terms of content. We removed seven items that measured other substantive variables (e.g. job satisfaction and affective organisational commitment) and 61 items that were not tied to the five work climate factors discussed in the text (i.e. one-item measures that were designed to assess employees' appreciation of a wide diversity of company-related policies and practices [e.g. training and development plans, target-setting and performance reviews, the job posting system, etc.], departmental goals). 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 operationalise the construct. Therefore, we conducted an exploratory factor analysis (EFA) on the remaining 21 items as precursor to the CFA (Gerbing & Hamilton, 1996; Hurley, Scandura, Schriesheim, Brannick, Seers, Vandenberg, & Williams, 1997). The total sample (with data from all 16 countries) was randomly split into two equally sized subsamples. Principal axis factoring (SPSS for Windows 14) followed by an oblique rotation method (orthogonal factors were not expected) was conducted using the data from the first subsample. Based on the "eigenvalue-greater-than-one" criterion we selected five factors and 13 items measuring these factors (i.e. they all had factor loadings exceeding .50; see Hair, Black, Babin, Anderson, & Tatham, 2006). The five-factor structure was confirmed (i.e. showing identical patterns of factor loadings) using the data from the other subsample.

Second, results of a previous wave of the climate survey (see Discussion) showed that the five factors under study show substantial correlations (ranging from .40 to .61) with a measure of key importance, namely a composite measure of job satisfaction (alpha value = .82).

ANALYSES

Sequence of Models Tested

We used structural equation modeling in the form of MACS models to test for measurement equivalence across modes of data collection. Mplus 2 (Muthén and Muthén, 1999, 2003) was used to evaluate all MACS models.

We started by testing a confirmatory factor analysis model which imposed the hypothesised five-factor structure (i.e. the key factors of the organisation's business model) onto the data. Strictly speaking, this model is not a MACS model as indicator mean scores are not needed to test the hypothesised dimensionality of the factor model. The five-factor model was evaluated using the samples from all individual countries.

Provided that the five-factor model fit the data well, a hierarchical sequence of nested statistical models (e.g. Vandenberg & Lance, 2000) can be used to assess measurement equivalence of indicator variables across modes of data collection. In this study, we investigated whether Meredith's (1993) 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 (modes of data collection). Unique variances of indicators (i.e. unreliabilities) may, however, vary across the alternative modes of data collection. The same is true for factor means, factor variances, and factor covariances. Meredith's equivalence condition is referred to as "scalar equivalence" across groups.

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, were constrained to be equal across modes of data collection. Conceptually, the baseline model assumes that the data exhibit (factor) "form equivalence" across modes of data collection. 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 regardless of the mode of data collection used. Conceptually, this means that employees use a similar frame-of-reference when completing the items of the organisational survey (Riordan & Vandenberg, 1994). An additional constraint in our baseline model is that

all observed variables load on just one factor (i.e. cross-loadings are not specified).

The second model in the sequence constrains all factor loadings to be identical across modes of data collection 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". Conceptually, equivalence of factor loadings implies that respondents calibrate the intervals used on the measurement scale in similar ways (Riordan & Vandenberg, 1994).

The third model in the sequence, the scalar equivalence model, constrains all factor loadings and indicator intercepts[3] to be identical across modes of data collection. The remaining parameters (i.e. unique variances, factor means, and factor variances and covariances) are not constrained across modes of data collection. 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 modes of data collection (Drasgow, 1984, 1987). Non-equivalence of indicator intercepts across modes of data collection may be caused by differences in "yea saying" (i.e. agreement bias, also known as "acquiescence response style bias") across these modes of data collection. It may well be that some data collection methods invoke a higher frequency of positive responses to items, regardless of item content (i.e. acquiescence bias).

Assessment of Model Fit

Apart from traditional fit indices (Chi-square statistic and Chi-square/df), we relied on four measures of model fit. 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-Bonett Non-Normed Fit Index (NNFI; Bentler & Bonett, 1980), (3) the Root Mean Square Error of Approximation (RMSEA; Steiger, 1990), and (4) the Standardized Root Mean Square Residual (SRMR; Bentler, 1995). These goodness-of-fit measures were suggested by the extensive simulation study of Hu and Bentler (1999). Hu and Bentler also proposed the following cutoff values: .95 (i.e. minimum values for CFI and TLI), .08 and .06 (i.e. maximum values for SRMR and RMSEA, respectively).

To statistically compare alternative measurement equivalence models we did not use the traditional Chi-square difference statistic. As this statistic is

[3] As indicated by one of the reviewers, when variable means are unequal across groups, constraining intercept differences to be equal across groups leads to factor mean differences, and conversely, constraining factor means to be equal will lead to differences in the indicator intercepts.

sensitive to sample size (Brannick, 1995; Kelloway, 1995), we used the difference in Comparative Fit Index (CFI) between nested equivalence models because a recent simulation study (Cheung & Rensvold, 2002) found that the difference in Comparative Fit Index was a reliable (and robust) measure of model fit (in measurement equivalence testing using MACS models). Specifically, the difference in CFI between (successive) equivalence models might not be higher than .01.

RESULTS

Test of the Baseline Five-Factor Model

Confirmatory factor analyses were run to test the adequacy of the five-factor structure using the data from each individual country. From Table 2 it may be concluded that, in general, the five-factor model provided an adequate representation of the data in all countries. For some countries, the TLI measure reported fell somewhat below the critical level of .95. The RMSEA measure was slightly too high in Sweden (.062) and in the United States (.064). As no other problems were encountered, we decided to proceed with an assessment of measurement equivalence across Internet and PP surveys in all (i.e. 16) countries.

TABLE 2
SEM Models Specifying a Five-factor Structure

Country	*Chi*^{2} *(d.f.)*	*CFI*	*TLI*	*RMSEA*	*SRMR*
Australia	331.4 (55)	.970	.957	.060	.043
Brazil	1,798.5 (55)	.966	.952	.057	.045
PR China	579.3 (55)	.969	.956	.058	.043
Czech Republic	252.7 (55)	.961	.945	.059	.050
France	569.0 (55)	.965	.950	.055	.042
Germany	1,081.6 (55)	.962	.946	.055	.039
The Netherlands	617.5 (55)	.966	.952	.052	.039
Nigeria	195.4 (55)	.964	.950	.051	.044
Pakistan	188.5 (55)	.968	.955	.054	.041
Puerto Rico	144.2 (55)	.975	.965	.051	.036
Russian Federation	314.1 (55)	.960	.943	.060	.041
Spain	327.7 (55)	.969	.955	.059	.044
Sweden	247.7 (55)	.964	.949	.062	.037
United Kingdom	1,612.2 (55)	.967	.954	.060	.037
United States	2,096.7 (55)	.965	.950	.064	.040
Vietnam	276.7 (55)	.965	.950	.054	.047

Measurement Equivalence of PP and Internet across Countries

Our hypothesis posited that, in every country, measurement equivalence would be established across Internet and PP surveys. Table 3 shows that the assumption of form equivalence of the constructs measured was supported for the majority of countries as all model fit values (except for, occasionally, the TLI measure) did not exceed the critical values of Hu and Bentler (1999). In some cases (i.e. Australia, Czech Republic, Sweden, and the US), the RMSEA measure was somewhat too high (RMSEA = .062 in Australia and Czech Republic and RMSEA = .065 in Sweden and the US). In addition, the RMSEA measure was rather high in the Russian Federation (.067).

As indicated in Table 3, the assumption of metric equivalence was supported in all countries. The difference in CFI between the form equivalence model and the metric model is consistently below the critical difference proposed by Cheung and Rensvold (2002) (i.e. CFI should not drop by more than .01). As a result, one may conclude that Internet and PP respondents calibrate the intervals used on the measurement scale in a similar way (see Riordan & Vandenberg, 1994).

The next model comparison involves a statistical comparison between the metric equivalence model and the scalar equivalence model. Based on the differences in CFI reported, the scalar equivalence model is acceptable in all countries except the People's Republic of China and France. However, the actual difference in CFI (.011) is only slightly above the critical difference of .01 in both countries. We therefore conclude that, in general, scalar equivalence is supported across the two modes of data collection which were combined in this mixed-mode organisational survey.

Measurement Equivalence of PP and Internet across Countries while Controlling for Job Level

Differences in job level may partially account for differences in response behavior across modes of data collection (see Roberts et al., 2004; Stanton & Rogelberg, 2001). Employees with a higher job level may show a stronger tendency to respond via the Internet than employees with a lower job level as the former group tends more often to have direct access to the Internet (and/or the Intranet) than the latter group. The data (aggregated across countries) show that employees with the highest and intermediate job level, in particular, respond more frequently via the Internet. The percentages of Internet responses are 66.4 per cent, 76.4 per cent, and 26.7 per cent within the highest, the intermediate and the lowest job levels, respectively. Because of the existing differences in Internet responding behavior, the earlier analyses were repeated using *matched subsamples* of Internet and PP surveys within each country.

TABLE 3
Within-Country Tests of Measurement Equivalence between Internet and Mail Surveys

Comparisons between Internet and PP surveys	(N_1; N_2) [N1/(N1 + N2)]	*Measurement equivalence model*	*Chi² (d.f.)*	*CFI*	*Difference in CFI (M1–M2 or M2–M3)*	*TLI*	*RMSEA*	*SRMR*
Country								
Australia		M1: Form	406.1 (110)	.966	—	.952	.062	.045
	(733; 676)	M2: Metric	428.6 (118)	.965	.001	.953	.061	.047
	[0.520]	M3: Scalar	4,66.7 (126)	.961	.004	.952	.062	.048
Brazil		M1: Form	1,910.4 (110)	.965	—	.951	.058	.046
	(1,236; 8,512)	M2: Metric	1,939.0 (118)	.965	.000	.953	.056	.046
	[0.127]	M3: Scalar	2,301.2 (126)	.958	.007	.948	.060	.048
PR China		M1: Form	629.5 (110)	.969	—	.956	.057	.044
	(957; 1,909)	M2: Metric	668.2 (118)	.967	.002	.957	.057	.047
	[0.334]	M3: Scalar	861.3 (126)	.956	.011	.946	.064	.050
Czech Republic		M1: Form	326.4 (110)	.958	—	.941	.062	.054
	(238; 780)	M2: Metric	362.0 (118)	.953	.005	.938	.064	.059
	[0.234]	M3: Scalar	411.7 (126)	.945	.008	.932	.067	.061
France		M1: Form	653.1 (110)	.963	—	.948	.056	.044
	(1,490; 1,645)	M2: Metric	682.2 (118)	.962	.001	.949	.055	.046
	[0.475]	M3: Scalar	839.5 (126)	.951	.011	.940	.060	.048
Germany		M1: Form	1,185.2 (110)	.961	—	.941	.056	.041
	(986; 5,235)	M2: Metric	1,219.3 (118)	.960	.001	.947	.055	.041
	[0.158]	M3: Scalar	1,361.0 (126)	.955	.005	.944	.056	.043
Netherlands		M1: Form	676.8 (110)	.966	—	.951	.052	.039
	(2,238; 1,528)	M2: Metric	693.4 (118)	.965	.001	.954	.051	.040
	[0.594]	M3: Scalar	827.2 (126)	.958	.007	.947	.054	.043
Nigeria		M1: Form	243.1 (110)	.967	—	.953	.049	.046
	(188; 801)	M2: Metric	252.9 (118)	.966	.001	.956	.048	.048
	[0.190]	M3: Scalar	304.8 (126)	.956	.010	.945	.054	.052

TABLE 3
Continued

Comparisons between Internet and PP surveys	*(N_1; N_2) [N1/(N1 + N2)]*	*Measurement equivalence model*	*Chi² (d.f.)*	*CFI*	*Difference in CFI (M1–M2 or M2–M3)*	*TLI*	*RMSEA*	*SRMR*
Pakistan		M1: Form	257.0 (110)	.966	—	.951	.057	.045
	(394; 432)	M2: Metric	273.1 (118)	.964	.002	.952	.056	.047
	[0.477]	M3: Scalar	290.0 (126)	.962	.002	.952	.056	.048
Puerto Rico		M1: Form	203.5 (110)	.975	—	.965	.052	.041
	(93; 541)	M2: Metric	218.2 (118)	.973	.002	.965	.052	.044
	[0.147]	M3: Scalar	246.9 (126)	.968	.005	.960	.055	.047
Russian Federation		M1: Form	431.8 (110)	.951	—	.930	.067	.047
	(672; 613)	M2: Metric	447.3 (118)	.950	.001	.933	.066	.049
	[0.523]	M3: Scalar	488.8 (126)	.945	.005	.931	.067	.050
Spain		M1: Form	392.2 (110)	.967	—	.953	.060	.045
	(547; 884)	M2: Metric	406.7 (118)	.967	.000	.956	.058	.047
	[0.382]	M3: Scalar	453.8 (126)	.962	.005	.953	.060	.049
Sweden		M1: Form	327.0 (110)	.959	—	.942	.065	.041
	(400; 525)	M2: Metric	342.0 (118)	.958	.001	.944	.064	.044
	[0.432]	M3: Scalar	368.4 (126)	.954	.004	.943	.064	.045
United Kingdom		M1: Form	1,662.7 (110)	.967	—	.953	.060	.037
	(3,279; 4,519)	M2: Metric	1,736.2 (118)	.965	.002	.954	.059	.041
	[0.420]	M3: Scalar	1,998.5 (126)	.960	.005	.950	.062	.043
United States		M1: Form	2,212.5 (110)	.964	—	.949	.065	.041
	(5,439; 3,593)	M2: Metric	2,309.3 (118)	.962	.002	.950	.064	.043
	[0.602]	M3: Scalar	2,625.5 (126)	.957	.005	.947	.066	.046
Vietnam		M1: Form	319.3 (110)	.966	—	.952	.053	.049
	(126; 1,252)	M2: Metric	329.7 (118)	.966	.000	.955	.051	.050
	[0.091]	M3: Scalar	369.7 (126)	.961	.005	.951	.053	.051

Notes: N_1 = number of Internet surveys; N_2 = number of mail surveys.

The following procedure was used. After inspection of the distribution of the variable job level in the Internet and PP sample (per country), the sample (either the Internet or the PP sample) containing the smallest number of employees with a given job level was identified. Next, using the other sample, a random subsample was drawn containing an identical number of employees holding that particular job level. This procedure was applied for all three job levels (see description in Sample section). Upon merging all of these (sub)samples, matched Internet and PP samples were obtained per country. Obviously, this procedure leads to smaller samples (ranging from .229 [Brazil] to .812 [Russian Federation] times the original sample size). The results of these additional analyses using matched subsamples are presented in Table 4. As shown in Table 4, the sample size of the matched Internet and PP samples in three countries (i.e. PR China, Puerto Rico, and Vietnam) was too small (i.e. $N < 90$) to repeat our earlier analyses. As a consequence, the stability of our previous research could only be tested for 13 countries ($N = 12{,}383$).

These additional analyses led to virtually the same overall conclusion, namely strong (cross-country) support for measurement equivalence (i.e. scalar equivalence) across Internet and PP surveys. In this matched-samples approach, there was even empirical support for scalar equivalence of Internet and PP surveys in France.

To allow the reader to have a closer look at the measurement parameters of the five-factor model which was selected as our final measurement model (i.e. the scalar equivalence model), two supplementary tables are provided. Table 5 provides the factor intercorrelations as estimated in the scalar equivalence model using all individual-level observations from all countries ($N = 52{,}461$). As shown in Table 5, the largest intercorrelations were found between F4 (Decision Making) and some other factors (e.g. with F2: Supervisor Support, $r = .66$; with F1: Team Commitment, $r = .53$; and with F3: Goal Clarity, $r = .50$). Table 6 presents the factor loadings and indicator intercepts as estimated in the scalar equivalence model. In addition, Table 6 provides details on the amount of variance in the indicator scores (i.e. R^2), which is explained by the factor they are supposed to measure.

DISCUSSION

This study was the first to conduct multiple country-specific analyses on the measurement equivalence of an organisational survey across Internet and PP survey administrations. This meant that we tested the equivalence of PP and Internet surveys in a wide variety of countries. Prior studies based their analyses either on data from just one country (i.e. mainly the United States or Spain) or non-homogeneous samples (aggregated across multiple countries). This study offered more rigorous (i.e. country-specific) tests on measurement

TABLE 4
Within-Country Tests of Measurement Equivalence between Internet and Mail Surveys (after Controlling for Job Level)

Comparisons between Internet and PP surveys	*(N)*	*Measurement equivalence model*	*Chi² (d.f.)*	*CFI*	*Difference in CFI (M1–M2 or M2–M3)*	*TLI*	*RMSEA*	*SRMR*
Country								
Australia		M1: Form	310.7 (110)	.970	—	.958	.060	.042
		M2: Metric	320.0 (118)	.970	.000	.960	.058	.044
	(547)	M3: Scalar	347.0 (126)	.967	.003	.959	.058	.045
Brazil		M1: Form	501.6 (110)	.967	—	.953	.058	.044
		M2: Metric	517.5 (118)	.966	.001	.955	.056	.045
	(1,117)	M3: Scalar	637.5 (126)	.957	.009	.946	.061	.049
Czech Republic		M1: Form	209.4 (110)	.943	—	.920	.069	.069
		M2: Metric	231.0 (118)	.936	.007	.915	.071	.077
	(197)	M3: Scalar	268.6 (126)	.919	.017	.899	.078	.080
France		M1: Form	384.4 (110)	.966	—	.952	.054	.044
		M2: Metric	399.6 (118)	.965	.001	.954	.053	.046
	(912)	M3: Scalar	462.9 (126)	.959	.006	.949	.056	.047
Germany		M1: Form	533.1 (110)	.949	—	.928	.066	.050
		M2: Metric	560.7 (118)	.947	.002	.929	.065	.052
	(935)	M3: Scalar	633.6 (126)	.939	.008	.924	.067	.054
Netherlands		M1: Form	557.3 (110)	.964	—	.949	.054	.041
		M2: Metric	571.1 (118)	.964	.000	.952	.052	.043
	(1,455)	M3: Scalar	696.6 (126)	.954	.010	.944	.057	.046
Nigeria		M1: Form	162.9 (110)	.960	—	.944	.058	.053
		M2: Metric	170.6 (118)	.960	.000	.948	.056	.059
	(176)	M3: Scalar	187.8 (126)	.954	.006	.942	.059	.062
Pakistan		M1: Form	233.2 (110)	.955	—	.936	.067	.049
		M2: Metric	257.1 (118)	.949	.006	.932	.068	.058
	(291)	M3: Scalar	265.1 (126)	.949	.000	.936	.066	.059

TABLE 4
Continued

Comparisons between Internet and PP surveys	*(N)*	*Measurement equivalence model*	*Chi² (d.f.)*	*CFI*	*Difference in CFI (M1–M2 or M2–M3)*	*TLI*	*RMSEA*	*SRMR*
Russian Federation		M1: Form	377.7 (110)	.948	—	.927	.069	.049
		M2: Metric	398.5 (118)	.946	.002	.928	.068	.051
	(522)	M3: Scalar	439.4 (126)	.939	.007	.925	.070	.053
Spain		M1: Form	299.1 (110)	.964	—	.949	.064	.046
		M2: Metric	306.2 (118)	.964	.000	.953	.062	.048
	(432)	M3: Scalar	348.4 (126)	.958	.006	.948	.065	.051
Sweden		M1: Form	309.9 (110)	.940	—	.916	.081	.051
		M2: Metric	324.9 (118)	.938	.002	.919	.079	.055
	(286)	M3: Scalar	336.9 (126)	.937	.001	.922	.077	.056
United Kingdom		M1: Form	999.1 (110)	.965	—	.951	.062	.039
		M2: Metric	1,020.9 (118)	.965	.000	.953	.060	.042
	(2,202)	M3: Scalar	1,158.4 (126)	.960	.005	.950	.062	.043
United States		M1: Form	1,597.5 (110)	.963	—	.947	.066	.041
		M2: Metric	1,655.6 (118)	.961	.002	.949	.064	.044
	(3,311)	M3: Scalar	1,781.1 (126)	.958	.003	.949	.065	.045

Notes: N = Sample size per group (Internet and mail); for countries with $N < 90$ no test statistics are reported (i.e. Puerto Rico, PR China, Vietnam).

TABLE 5
Factor Intercorrelations as Estimated in the Scalar Equivalence Model[1]

	F1	*F2*	*F3*	*F4*	*F5*
F1	1.00				
F2	.52	1.00			
F3	.35	.40	1.00		
F4	.53	.66	.50	1.00	
F5	.24	.25	.28	.38	1.00

Note: [1] All individual-level observations from all 16 countries were used in this analysis (N = 52,461).

equivalence of a global survey instrument. In addition, we used the more advanced and stringent MACS approach to measurement equivalence testing.

The most important result of this study is that we found evidence for scalar equivalence of the multi-item survey instrument across Internet and PP surveys virtually in all countries in which the organisation survey was administered. These findings are good news for international survey researchers as they provide an empirical justification for using, combining, and comparing data from mixed-mode surveys in various countries. In addition, across-country (collapsing all data of all countries) MACS results corroborated the within-country results as scalar invariance was found between Internet and PP surveys (results are available from the first author). Next, additional analyses also showed that, across both modes of data collection, high rank correlations (i.e. most often above .80) were obtained between the mean indicator scores of all five factors included in the study. This is an indication that, in an applied context, both modes of data collection can be combined without substantially affecting the results of the study.

This study also investigated whether job level served as moderator of measurement equivalence between Internet and PP survey data. However, measurement equivalence results were virtually similar in our original analyses as in our analyses wherein we controlled for job level. This suggests that job level did not play a critical role.

Some limitations should be acknowledged. First, this study is based on a premise that is common in cross-cultural psychology, namely that the country is the unit of analysis. Yet, we acknowledge that there are also cultural and linguistic variations within a given country. However, from an applied administration point of view, countries form very important entities in multinational companies. So, the focus on countries (rather than on cultures) is justifiable. Second, some might argue that the organisation-specific survey under investigation in this study was not an established measure. We

TABLE 6
Factor Loadings and Indicator Intercepts as Estimated in the Scalar Equivalence Model

	Scalar Equivalence Model[1,2]	*Indicator Intercept*	SE	t-*value*	*Unstandardised factor loading*	SE	t-*value*	R^2 *(percentage of variance explained)*
F1	Item 1[#]	2.380	.005	448.9	1.000	.000	—	.521
	Item 2	2.246	.005	449.0	.848	.007	117.0	.420
	Item 3	2.301	.005	447.8	.909	.008	120.2	.458
F2	Item 1[#]	2.515	.006	405.9	1.000	.000	—	.517
	Item 2	2.183	.006	385.4	1.05	.006	164.1	.689
	Item 3	2.204	.006	391.6	1.02	.006	161.7	.653
F3	Item 1[#]	1.657	.004	408.3	1.000	.000	—	.760
	Item 2	1.842	.004	411.6	1.13	.005	211.4	.798
	Item 3	2.129	.005	421.8	.705	.006	113.0	.504
F4	Item 1[#]	2.374	.005	439.6	1.000	.000	—	.847
	Item 2	2.504	.005	469.6	.830	.005	164.6	.599
F5	Item 1[#]	1.637	.004	422.3	1.000	.000	—	.580
	Item 2	1.743	.004	423.1	1.187	.010	120.4	.724

Notes: [1] Model fit statistics: $Chi^2 = 13{,}572.5$, d.f. = 82, CFI = .954, TLI = .941, RMSEA = .057; [2] All individual-level observations from the 16 countries ($N = 52{,}461$) were used in this analysis; [#] This indicator served as the reference indicator in the analysis; item numbers are in the same order as the items listed in the Appendix.

do not see this as a serious limitation. Consistent with Ryan et al. (1999), we believe that the essence of organisational surveying is that customised measures are constructed that enable an organisation to achieve its specific purposes. Nevertheless, it should be clear that future studies are needed to confirm our results in other organisations, in other countries, and with other measures. Third, as this was a quasi-experiment, participants were not randomly assigned to data collection modes as the choice of data collection method was based on personal or organisational grounds. This practice mirrored how the mixed-mode data collection approach is used in organisations. Practical constraints in data collection also prevented us from obtaining the response rates per country and data collection. Finally, we also concentrated on only two modes of data collection. Clearly, other modes of data collection are possible such as telephone-based or pda-based data collection. Future studies are needed to examine measurement equivalence of organisational survey data across these data collection formats.

In conlusion, the key message of this study is that it provided an empirical justification for the combined use of Internet and PP surveys in international organisational surveys. This is excellent news for the research community as mixed-mode surveys are growing in popularity in this particular area of research.

APPENDIX

Measures of the Five Organisational Climate Factors

Factor 1: Team Commitment ($\alpha = .72$[b])

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

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

Factor 2: Supervisor Support ($\alpha = .82$[b])

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

- 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 ($\alpha = .76$[b])

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 organisation.

- 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 organisation.
- I have a clear understanding of the goals and objectives of the multinational as a whole.

Factor 4: Decision Making ($\alpha = .83$[b])

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

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

Factor 5: Environmental and Societal Responsibility ($\alpha = .79$[b])

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

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

Notes: [a] This item was arbitrarily chosen as reference indicator in the analyses.
[b] The calculation is based on all observations (i.e. all country samples merged together).

REFERENCES

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

Bentler, P.M. (1995). *EQS structural equations program manual*. Encino, CA: Multivariate Software.

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

Birnbaum, M.H. (2001). *Introduction to behavioral research on the internet*. Upper Saddle River, NJ: Prentice Hall.

Booth-Kewley, S., Edwards, J.E., & Rosenfeld, P. (1992). Impression management, social desirability, and computer administration of attitude questionnaires: Does this make a difference? *Journal of Applied Psychology*, *77*, 562–566.

Bowling, A. (2005). Mode of questionnaire administration can have serious effects on data quality. *Journal of Public Health*, *27*, 281–291.

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

Buchanan, T., & Smith, J.L. (1999). Using the internet for psychological research: Personality testing on the World Wide Web. *British Journal of Psychology*, *90*, 125–144.

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*, 169–199.

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

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

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.

Cook, C., Heath, F., Thompson, R.L., & Thompson, B. (2001). Score reliability in Web-or-Internet-based surveys: Unnumbered graphic rating scales versus Likert-type scales. *Educational and Psychological Measurement*, *61*, 697–706.

Dillman, D. (2000). *Mail and internet surveys: The tailored design method* (2nd edn.). New York: J. Wiley.

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.

Epstein, J., Klinkenberg, W.D., Wiley, D., & McKinley, L. (2001). Ensuring sampling equivalence across Internet and paper-and-pencil assessments. *Computers in Human Behavior*, *17*, 339–346.

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 Francisco, CA: Jossey-Bass.

Ferrando, P.J., & Lorenzo-Seva, U. (2005). IRT-based factor analytic procedures for testing the equivalence of paper-and-pencil and internet-administered questionnaires. *Psychological Methods*, *10*, 193–205.

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.

Hair J.F., Jr., Black, W.C., Babin, B.J., Anderson, R.E., & Tatham, R.L. (2006). *Multivariate data analysis* (6th edn.). Upper Saddle River, NJ: Pearson Prentice Hall.

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*, 1–55.

Hurley, A.E., Scandura, T.A., Schriesheim, C.A., Brannick, M.T., Seers, A., Vandenberg, R.J., & Williams, L.J. (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, CA: 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, CA: Jossey-Bass.

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

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

Kopelman, R.E., Brief, A.P., & Guzzo, R.A. (1990). The role of climate and culture in productivity. In B. Schneider (Ed.), *Organisational climate and culture* (pp. 282–318). San Francisco, CA: Jossey-Bass.

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

Kraut, A.I., & Saari, L.M. (1999). Organizational surveys: Coming of age for a new era. In A.I. Kraut & A.K. Korman (Eds.), *Evolving practices in human resource management* (pp. 302–327). San Francisco, CA: Jossey-Bass.

Lautenschlager, G.J., & Flaherty, V.L. (1990). Computer administration of questions: More desirable or more social desirability? *Journal of Applied Psychology*, *75*, 310–314.

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

Mann, C., & Stewart, F. (2000). *Internet communication and qualitative research: A handbook for researching online*. London: Sage.

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

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.

Ostroff, C. (1993). The effects of climate and personal influences on individual behavior and attitudes in organizations. *Organizational Behavior and Human Decision Processes*, *56*, 56–90.

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

Ployhart, R.E., Weekley, J.A., Holtz, B.C., & Kemp, C. (2003). Web-based and paper-and-pencil testing of applicants in a proctored setting: Are personality, biodata, and situational judgment tests comparable? *Personnel Psychology*, *56*, 733–752.

Reips, U.D. (2000). The web experiment method: Advantages, disadvantages and solutions. In M.H. Birnbaum (Ed.), *Psychological experiments on the internet* (pp. 89–117). San Diego, CA: Academic Press.

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

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.

Schaeffer, D., & Dillman, D. (1998). Development of a standard e-mail methodology: Results of an experiment. *Public Opinion Quarterly*, *62*, 378–397.

Simsek, Z., & Veigha, J.F. (2001). A primer on internet organizational surveys. *Organizational Research Methods*, *4*, 218–235.

Sörbom, D. (1974). A general method for studying differences in factor means and factor structure across 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.

Sproull, L.S. (1986). Using electronic mail for data collection in organizational research. *Academy of Management Journal*, *29*, 159–169.

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

Stanton, J.M., & Rogelberg, S.G. (2001). Using internet/intranet web pages to collect organizational research data. *Organizational Research Methods*, *4*, 200–217.

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

Thompson, L.F., Surface, E.A., Martin, D.L., & Sanders, M.G. (2003). From paper to pixels: Moving personnel surveys to the web. *Personnel Psychology*, *56*, 197–227.

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 invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*, 4–70.

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.

Van Selm, M., & Jankowski, N.W. (2006). Conducting online surveys. *Quality & Quantity*, *40*, 435–456.

Vispoel, W.P., Boo, J., & Bleiler, T. (2001). Computerized and paper-and-pencil versions of the Rosenberg self-esteem scale: A comparison of psychometric features and respondent preferences. *Educational and Psychological Measurement*, *61*, 461–474.

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

Yun, G.W., & Trumbo, C.W. (2000). Comparative response to a survey executed by post, email and web form. *Journal of Computer-Mediated Communication*, *6*, 1–26.