Journal of International Business Studies (2008) 39, 1027–1044

www.jibs.net

Data equivalence in cross-cultural international business research: assessment and guidelines

G Tomas M Hult[1], David J Ketchen Jr[2], David A Griffith[1], Carol A Finnegan[3], Tracy Gonzalez-Padron[1], Nukhet Harmancioglu[4], Ying Huang[5], M Berk Talay[6] and S Tamer Cavusgil[1]

[1]Eli Broad Graduate School of Management, Michigan State University, East Lansing, USA; [2]Auburn University, Auburn, USA; [3]University of Colorado, Colorado Springs, USA; [4]Bilkent University, Ankara, Turkey; [5]University of Arizona, Tucson, USA; [6]HEC Montréal, Quebec, Canada

Correspondence:
GTM Hult, Michigan State University, Eli Broad Graduate School of Management, East Lansing, MI 48824-1121, USA.
Tel: +1 517 353 4336;
Fax: +1 517 432 1009;
E-mail: hult@msu.edu

Received: 23 May 2005
Revised: 12 December 2006
Accepted: 21 August 2007
Online publication date: 22 May 2008

Abstract
Data equivalence refers to the extent to which the elements of a research design have the same meaning, and can be applied in the same way, in different cultural contexts. Failure to establish data equivalence in cross-cultural studies may bias empirical results and theoretical inferences. Although several authors have encouraged researchers to ensure high levels of data equivalence, no study has assessed the status of the field in relation to compliance with data equivalence standards. Accordingly, this study examines three aspects of data equivalence (construct equivalence, measurement equivalence, and data collection equivalence) within 167 studies that involve cross-cultural data published in the *Journal of International Business Studies*, *Management International Review*, *Journal of World Business*, *Strategic Management Journal* and the *Academy of Management Journal* from 1995 to 2005. The findings indicate that international business researchers report insufficient information in relation to data equivalence issues, thus limiting confidence in the findings of many cross-cultural studies. To enhance future research, a guideline for procedures for researchers to follow and report in establishing data equivalence is offered.
Journal of International Business Studies (2008) **39,** 1027–1044.
doi:10.1057/palgrave.jibs.8400396

Keywords: data equivalence; construct equivalence; measurement equivalence; data collection equivalence

INTRODUCTION

Establishing the extent to which concepts, theories, and findings developed in one culture apply in other cultures has long been a primary research area within the field of international business (IB) research (e. g., Sekaran, 1983; Wright, 1970). IB researchers conduct cross-cultural studies for two main purposes: (1) to provide evidence addressing the generalizability of implications across borders; and (2) to understand any culture-specific differences regarding phenomena and relationships (Mintu, Calantone, & Gassenheimer, 1994). Further, cross-cultural studies are becoming increasingly important for research inquiry, teaching, and practice in functional business disciplines. However, for these studies to contribute to the field of IB research, a number of unique challenges must be overcome.

One key challenge is that, because of cultural differences, elements of research designs (such as survey items) cannot simply be exchanged in original form between cultures. For example, Japanese survey respondents tend to cluster their answers near the

center points of scales, as a function of their culture's emphasis on avoiding extreme positions (cf. Steenkamp & Baumgartner, 1998). As a result, a scale assessing individuals' stress levels that generates wide variance in responses in the US would be likely to produce less variance in Japan, thus possibly leading one to conclude that Americans vary more in their stress levels than do Japanese. However, the differences found may merely be a function of scale use, and thereby not reflect any "actual" cultural differences in relation to stress levels. This example illustrates the importance of creating *data equivalence* – taking steps to ensure that any differences found between cultures truly reflect the phenomena of interest, and are not simply a reflection of issues such as scale use tendencies and differences in construct conceptualizations.

In methodological terms, the failure to establish data equivalence is a potential source of measurement error (i.e., discrepancies between what is intended to be measured and what is actually measured), which attenuates the precision of estimators, reduces the power of statistical tests of hypotheses, and provides misleading results (Davis, Douglas, & Silk, 1981; van de Vijver & Leung, 1997). Although it is widely acknowledged that data equivalence issues have significant consequences for the findings of individual studies, and for knowledge generation in the IB field as a whole (e.g., Craig & Douglas, 2000; Mullen, 1995; Singh, 1995), it is unclear as to the extent to which current cross-cultural IB research assesses data equivalence issues and thereby avoids certain threats to reliability and validity. The purpose of this study is therefore to assess the current cross-cultural IB literature in relation to data equivalence. Further, our purpose is to build upon this analysis to offer recommendations for the field.

Accordingly, we examined the treatment of data equivalence within 167 IB studies published in *Journal of International Business Studies*, *Management International Review*, *Journal of World Business*, *Strategic Management Journal* and the *Academy of Management Journal* between 1995 and 2005. Specifically, we examine the data equivalence activities reported within these studies to the guidelines offered in the literature (e.g., Craig & Douglas, 2000; Mullen, 1995; Singh, 1995) for construct, measurement, and data collection equivalences. Here construct equivalence signifies whether a given concept or behavior has the same meaning and function from culture to culture (Kumar, 2000); measurement equivalence refers to the relative comparability of the wording, scaling, and scoring of constructs across cultures (Craig & Douglas, 2000; Mullen, 1995); and data collection equivalence relates to the comparability of sampling frames and the techniques used to gather data in each culture (Reynolds, Simintiras, & Diamantopoulos, 2003). The findings indicate that although some cross-cultural IB studies effectively establish data equivalence, other studies may take inadequate steps to establish data equivalence between cultures (as indicated by what is reported). Building on these findings, we provide suggestions and guidelines for future research that are intended to help the IB field maximize its potential for generating accurate insights into cross-cultural phenomena.

METHOD

The study focused on articles from 1995 to 2005 in order to consider how well cross-cultural IB researchers have dealt with data equivalence issues. The initial year of this study was chosen to be 1995, given the seminal studies that year in data equivalence measurement (i.e., Mullen, 1995; Singh, 1995). Articles were identified for inclusion in the study using the ABI Inform database with keyword searches including equivalence, data equivalence, measurement equivalence, construct equivalence, cross-cultural, cross-national, and cross-border. The sample was further restricted to the leading journals in the fields of IB (DuBois & Reeb, 2000) (i.e., *Journal of International Business Studies* (*JIBS*), *Management International Review* (*MIR*), and *Journal of World Business* (*JWB*)) and business strategy (Tahai & Meyer, 1999) (i.e., *Strategic Management Journal* (*SMJ*) and *Academy of Management Journal* (*AMJ*)), as these journals:

(1) are leading outlets for IB research;
(2) are purported to have the most rigorous research standards; and
(3) have the greatest impact on the field of IB, based upon citation rates.

Finally, to minimize research type confounds in the study, the sample excluded qualitative research, experimental studies, mathematical modeling papers, and studies that did not include data collection in more than one culture.[1] Table 1 lists the 167 articles that were included in our analysis. *JIBS* offered the most articles (79 articles, 47%), followed by *JWB* (34 articles, 20%), *MIR* (24 articles,

14%), *AMJ* (17 articles, 10%) and *SMJ* (13 articles, 8%).[2]

Two authors coded each article independently. To examine data equivalence, each article's treatment of construct, measurement, and data collection equivalence was coded (see Appendix). Each article was assessed in relation to data equivalence standards using statistical criteria from the extant IB and general business methods literature (e.g., Bollen, 1989; Craig & Douglas, 2000; Kumar, 2000; Mullen, 1995; Steenkamp & Baumgartner, 1998). Both coders had to agree that an article fitted the study for the article for it to be included in the analysis. Disagreements relating to the evaluation

Table 1 Articles included in the analysis

Academy of Management Journal (n=17)

Chen (1995)
Janssens et al. (1995)
Milliman et al. (1995)
Peterson et al. (1995)
Van De Vliert et al. (1996)
Gomez et al. (2000)
Lam and Schaubroeck (2000)
Mitchell et al. (2000)
Yang et al. (2000)
Kirkman and Shapiro (2001)
Spencer (2001)
Steensma et al. (2005)
Li and Hambrick (2005)
Takeuchi et al. (2005)
Van Der Vegt and Bunderson (2005)
Wang et al. (2005)
Luo (2005)

Management International Review (n=24)

Clarke and Hammer (1995)
Yavas (1995)
Mascarenhas and Sambharya (1996)
Ulgado (1996)
Buhner et al. (1997)
Hannon (1997)
Yip et al. (1997)
Beldona et al. (1998)
Robertson (2000)
Nachum (2001)
Stottinger and Holzmuller (2001)
Agarwal et al. (2002)
Glaum and Rinker (2002)
Lenartowicz and Johnson (2002)
Tan (2002)
Fahy et al. (2003)
Insch (2003)
Zhao et al. (2003)
Contractor et al. (2005)
Husted (2005)
Javidan and Carl (2005)
Kolk (2005)
Wang et al. (2005)
Zhao and Luo (2005)

Strategic Management Journal (n=13)

Geletkanycz (1997)
Very et al. (1997)
Gedajlovic and Shapiro (1998)
Bensaou *et al*. (1999)
Homburg et al. (1999)
Song et al. (1999)
Lane et al. (2001)
Subramaniam and Venkatraman (2001)
Luo (2002)
Brouthers et al. (2003)
Mayer and Whittington (2003)
Barr and Glynn (2004)
Hoskisson et al. (2004)

Journal of International Business Studies (n=79)

Cosset and Suret (1995)
Cullen et al. (1995)
Gibson (1995)
Salter and Niswander (1995)
Saudagaran and Biddle (1995)
Schlegelmilch and Robertson (1995)
Shane (1995)
Bigoness and Blakely (1996)
Dawar et al. (1996)
Husted et al. (1996)
Johnson et al. (1996)
Smith et al. (1996)
Bailey et al. (1997)
Barkema and Vermeulen (1997)
Dyer and Song (1997)
Kim and Chung (1997)
Ralston et al. (1997)
Zaheer and Zaheer (1997)
Kashlak et al. (1998)
Lin and Germain (1998)
Morris et al. (1998)
Agarwal et al. (1999)
Borkowski (1999)
Cadogan et al. (1999)
Pornpitakpan (1999)
Ralston et al. (1999)
Soutar et al. (1999)
Whitman et al. (1999)
Bowman et al. (2000)
Dyer and Chu (2000)
Fahy et al. (2000)
Griffith et al. (2000)
Harzing (2000)
Lee et al. (2000)
Neelankavil et al. (2000)
Steensma et al. (2000)
Veiga et al. (2000)
Balabanis et al. (2001)
Begley and Tan (2001)
Filatotchev et al. (2001)
Griffith and Harvey (2001)
Lau and Ngo (2001)
Lynch and Beck (2001)
Manev and Stevenson (2001)
Marshall and Boush (2001)
Chen and Hennart (2002)
Huang and Van de Vliert (2002)
Hofstede et al. (2002)
Skarmeas et al. (2002)
Thomas and Au (2002)
Giacobbe-Miller et al. (2003)
Lenartowicz and Johnson (2003)
Steenkamp et al. (2003)
Van de Vliert (2003)
Björkman et al. (2004)
Fock et al. (2004)
Fu et al. (2004)
Jensen and Szulanski (2004)
Shay and Baack (2004)
Beaulieu et al. (2005)
Belderbos and Heijltjes (2005)
Faff and Marshall (2005)
Hillman and Wan (2005)
Luo (2005)
Maitland et al. (2005)
Meschi (2005)
Murray et al. (2005)
Nachum and Wymbs (2005)
Oxelheim and Randøy (2005)
Rao et al. (2005)
Ruckman (2005)
Vaaler et al. (2005)

Table 1 *Continued*

Journal of International Business Studies (continued)		
Heuer et al. (1999)	Kwok (2002)	Venaik et al. (2005)
Husted (1999)	Maignan and Ralston (2002)	
Money and Graham (1999)	Makhija and Stewart (2002)	
Journal of World Business (n=34)		
Brewster et al. (1997)	Hult et al. (2000)	Jesuino (2002)
Sparrow and Budhwar (1997)	Kotabe et al. (2000)	Kabasakal and Bodur (2002)
Sim and Ali (1998)	Chan and Holbert (2001)	Peterson et al. (2002)
Tung (1998)	Robertson et al. (2001)	Ramus (2002)
Bjorkman and Lu (1999)	Shi (2001)	Szabo et al. (2002)
Child and Yan (1999)	Wong et al. (2001)	Wright et al. (2002)
Hegarty and Tihanyi (1999)	Ashkanasy et al. (2002)	Parboteeah et al. (2004)
Punnett and Clemens (1999)	Au and Fukuda (2002)	Roth et al. (2004)
Selmer (1999)	Bakacsi et al. (2002)	Haahti et al. (2005)
Wang et al. (1999)	Gupta et al. (2002)	Ralston et al. (2005)
Ang (2000)	Gupta et al. (2002)	
Cullen et al. (2000)	Harpaz et al. (2002)	

Note: Full details of the papers listed in this table are available from the authors on request.

of article elements were resolved through discussion. Inter-rater reliability was 85%, which is comparable to others' coding of published studies (e.g., Combs & Ketchen, 2003).

Diagnostics of the sample indicate that 138 of the studies (83%) were cross-sectional, 18 were longitudinal, and 11 relied on a lagged cross-sectional approach. Of the studies, 132 (79%) used primary data, of which 10 studies relied on interviews or personally administered questionnaires. Thirty studies used archival data (18%), and the remaining five studies used a variety of other data sources. Seventy-one studies (43%) focused on the individual customer as the level of analysis, and 50 studies (30%) adopted the organization as the key level. The remaining studies focused on levels such as groups, projects, and alliances. The most frequent number of countries examined was two (45 studies, 30%), followed by three (31 studies, 19%), and four (16 studies, 10%). Thirty-seven studies examined 10 or more countries. The most popular analytical technique used was regression (54 studies, 32%), followed by analysis of variance (38 studies, 23%). Other techniques represented included structural equation modeling, time series analysis, and hierarchical linear modeling.

CONSTRUCT EQUIVALENCE

Background

Construct equivalence relates to whether an object, concept, or behavior is the same (i.e., serves the same purpose and achieves the same salience) in all contexts and cultures (Craig & Douglas, 2000; Kumar, 2000). In the cross-cultural literature, both *etic* (i.e., universal) and *emic* (i.e., culturally specific) measures may be used to represent the theoretical domain of the construct fully and equally across cultures (Mintu et al., 1994). Kumar (2000) argues that rigorous research should be able to capture both commonalities (i.e., the *etic*) and uniqueness (i.e., the *emic*) in the meaning of constructs in and across cultures. Measures to protect construct equivalence are recommend to be engaged in by IB researchers both pre- and post-data collection.

Prior to data collection, research guidelines indicate that three types of construct equivalence are recommended in the research process: functional (i.e., the extent to which the objects and behavior take the same role or function across cultures); conceptual (i.e., the extent to which the domains of the concept/behavior are the same across cultures); and category (i.e., the extent to which the same classification scheme can be used for the concept and behavior across cultures) (cf. Craig & Douglas, 2000). Neglect of construct discrepancies often, if not always, leads to misleading results, given differences in the underlying constructs (Cavusgil & Das, 1997).

Further, a number of measures are available to IB researchers to assess construct equivalence post-data collection. Specifically, the most common tests of construct equivalence post-data collection are tests for unidimensionality (e.g., exploratory factor analysis (EFA) and confirmatory factor analysis

(CFA)), reliability (e.g., Cronbach's alpha and composite reliability; Cronbach, 1951; Fornell & Larcker, 1981; Nunnally, 1978), and construct validity (convergent and discriminant validity; Anderson & Gerbing, 1988; Bollen, 1989; Churchill, 1979; Fornell & Larcker, 1981; Nunnally, 1978). Articles were therefore assessed in relation to their reporting of unidimensionality, reliability, and construct validity (cf. Bollen, 1989; Campbell & Fiske, 1959; Gerbing & Anderson, 1992).

Findings

The findings relating to pre-data collection construct equivalence issues are reported in relation to functional, conceptual and category equivalence, and indicate a lack of pre-data collection construct equivalence (see Tables 2 and 3). Specifically, only 36% of the studies reported the establishment of functional equivalence (with 32% of studies presenting functional equivalence testing between 1995 and 1999, and 38% doing so between 2000 and 2005 (t=0.733, p=0.465)). The results indicate that although few studies have reported the establishment of functional equivalence, some strides have been made more recently to improve the reporting of this construct equivalence issue. An examination across journals found differences across journals (F=3.353, p=0.012) with *SMJ* reporting functional equivalence significantly more than *JWB* (p=0.018).[3] Further, *AMJ*, *SMJ*, and *MIR* were the only outlets to exceed the average in terms of functional equivalence reporting. In terms of conceptual equivalence, only 40% of the studies reported the establishment of conceptual equivalence (with 39% of studies addressing this type of equivalence between 1995 and 1999, and 42% doing so between 2000 and 2005 (t=0.321, p=0.748)). An examination across journals found statistical differences (F=2.498, p=0.045), with *SMJ* reporting conceptual equivalence significantly more than *JWB* (p=0.045). *AMJ*, *SMJ*, and *MIR* were the only outlets to exceed the average in reporting conceptual equivalence. Overall, only 19% of studies examined category equivalence (with the percentage of studies reporting category equivalence slightly declining from 20% to 18% over the two time periods (t=−0.362, p=0.718)). Further, although there were no statistical differences across journals (F=0.361, $p > 0.05$), *JIBS*, *MIR*, and *SMJ* all exceeded the 19% average (with the strongest performance by *SMJ*, with 25% of the 12 relevant studies examining category equivalence). The overall lack of construct equivalence reporting in relation to pre-data collection methods is concerning, as failure to establish functional, conceptual, or category equivalence threatens the validity and credibility of conclusions of IB research.

The findings relating to post-data collection construct equivalence issues, reported in relation to unidimensionality, reliability, and construct validity, are also concerning. Overall, unidimensionality was assessed in only 38% of the studies examined (with 37% of these studies reporting unidimensionality between 1995 and 1999, and 40% doing so between 2000 and 2005 (t=0.328, p=0.743)). As noted in Table 2, statistical differences were found across groups in relation to the reporting of unidimensionality (F=8.747, p=0.000), with *SMJ* being statistically different from *JWB* (p=0.001) and *MIR* (p=0.045) but not statistically different from *AMJ* or *JIBS*. Further, *JWB* was not found to be different from *MIR* (p=0.728), but *JWB* was statistically different from *JIBS* (p=0.012). *AMJ* and *SMJ* exceeded the average, with 77% reporting unidimensionality. Reliability was reported in only 55% of the studies examined (with 50% reporting reliability in the first time period and 58% in the second time period (t=0.987, p=0.325)). Again, there were differences across journals (F=3.533, p=0.009), with the only significant difference being between *AMJ* and *JWB* (p=0.042). Only *AMJ* and *SMJ* exceeded the average, with 77% reporting reliability. Construct validity was reported in only 41% of the studies examined (with 37% of the studies between 1995 and 1999 and 43% between 2000 and 2005 doing so (t=0.795, p=0.428)). As noted in Table 2, statistical differences were found across groups in relation to the reporting of construct validity (F=6.003, p=0.000), with *SMJ* not being statistically different from any of the other journals, but with *AMJ* being different from *JWB* (p=0.000) and *MIR* (p=0.020), and *JIBS* being different from *JWB* (p=0.004). Only *AMJ* and *SMJ* exceeded the average, with 77% reporting construct validity.

Implications

The findings in relation to the establishment of both pre- and post-data collection construct equivalence are concerning, because without cross-cultural construct equivalence any inferences made from empirical studies are subject to undermining threats. These findings are particularly disconcerting in light of IB's heavy reliance on perceptual measures. A significant number of perceptual measures originate in Western markets

Table 2 Treatment of data equivalence issues by journal

	F (sig.)	*Categories*	*AMJ*	*JIBS*	*JWB*	*MIR*	*SMJ*	*Total*
			(n=17)	*(n=79)*	*(n=34)*	*(n=24)*	*(n=13)*	*(n=167)*
Construct equivalence								
Functional equivalence	3.353 (p=0.012)	Reported	7 (46.7%)	25 (32.9%)	7 (20.6%)	9 (40.9%)	9 (75.0%)	57 (35.8%)
		Not reported	8 (53.3%)	51 (67.1%)	27 (79.4%)	13 (59.1%)	3 (25.0%)	102 (64.2%)
		NA	2	3	0	2	1	8
Conceptual equivalence	2.498 (p=0.045)	Reported	8 (50.0%)	29 (38.2%)	9 (26.5%)	10 (45.5%)	9 (75.0%)	65 (40.6%)
		Not reported	8 (50.0%)	47 (61.8%)	25 (73.5%)	12 (54.5%)	3 (25.0%)	95 (59.4%)
		NA	1	3	0	2	1	7
Category equivalence	0.361 (p=0.836)	Reported	2 (12.5%)	15 (19.7%)	5 (14.7%)	5 (23.8%)	3 (25.0%)	30 (18.9%)
		Not reported	14 (87.5%)	61 (80.3%)	29 (85.3%)	16 (76.2%)	9 (75.0%)	129 (81.1%)
		NA	1	3	0	3	1	8
Unidimensionality	8.747 (p=0.000)	Reported	13 (76.5%)	30 (38.0%)	4 (11.8%)	7 (29.2%)	10 (76.9%)	64 (38.3%)
		Not reported	4 (23.5%)	49 (62.0%)	30 (88.2%)	17 (70.8%)	3 (23.1%)	103 (61.7%)
Reliability	3.533 (p=0.009)	Reported	13 (76.5%)	47 (59.5%)	12 (35.3%)	10 (41.7%)	10 (76.9%)	92 (55.1%)
		Not reported	4 23.5%	32 40.5%	22 (64.7%)	14 (58.3%)	3 (23.1%)	75 (44.9%)
Construct validity	6.003 (p=0.000)	Reported	13 (76.5%)	36 (45.6%)	5 (14.7%)	7 (29.2%)	7 (53.8%)	68 (40.7%)
		Not reported	4 (23.5%)	43 (54.4%)	29 (85.3%)	17 (70.8%)	6 (46.2%)	99 (59.3%)
Measurement equivalence								
Calibration equivalence	1.403 (p=0.236)	Reported	3 (21.4%)	12 (16.7%)	2 (6.3%)	2 (11.8%)	4 (33.3%)	23 (15.6%)
		Not reported	11 (78.6%)	60 (83.3%)	30 (93.8%)	15 (88.2%)	8 (66.7%)	124 (84.4%)
		NA	3	7	2	7	1	20
Translation equivalence	3.638 (p=0.007)	Reported	13 (81.3%)	39 (53.4%)	9 (27.3%)	9 (45.0%)	6 (54.5%)	76 (49.7%)
		Not reported	3 (18.8%)	34 (46.6%)	24 (72.7%)	11 (55.0%)	5 (45.5%)	77 (50.3%)
		NA	1	6	1	4	2	14
Metric equivalence	3.22 (p=0.014)	Reported	6 (37.50%)	23 (30.67%)	3 8.82%	3 (14.29%)	6 (50.00%)	41 (25.95%)
		Not reported	10 (62.50%)	52 (69.33%)	31 (91.18%)	18 (85.71%)	6 (50.00%)	117 (74.05%)
		NA	1	4	0	3	1	9
Scoring consistency	2.065 (p=0.088)	Reported	6 (40.00%)	17 (22.67%)	3 8.82%	3 (14.29%)	4 (33.33%)	33 (20.89%)
		Not reported	9 (60.00%)	58 (77.33%)	31 (91.18%)	18 (85.71%)	8 (66.67%)	125 (79.11%)
		NA	2	4	0	3	1	10
Scalar equivalence	3.257 (p=0.014)	Reported	6 (40.00%)	17 (22.67%)	1 2.94%	2 9.52%	3 (25.00%)	29 (18.47%)
		Not reported	9 (60.00%)	58 (77.33%)	33 (97.06%)	19 (90.48%)	9 (75.00%)	128 (81.53%)
		NA	2	4	0	3	1	10
Data collection equivalence								
Sampling frame comparability	1.916 (p=0.110)	Reported	2 (11.76%)	24 (32.43%)	8 (23.53%)	2 9.52%	5 (38.46%)	41 (25.79%)
		Not reported	15 (88.24%)	50 (67.57%)	26 (76.47%)	19 (90.48%)	8 (61.54%)	118 (74.21%)
		NA	0	5	0	3	0	8
Data collection procedure	2.403 (p=0.053)	Reported	6 (35.29%)	38 (56.72%)	14 (43.75%)	8 (44.44%)	11 (84.62%)	77 (52.38%)
		Not reported	11 (64.71%)	29 (43.28%)	18 (56.25%)	10 (55.56%)	2 (15.38%)	70 (47.62%)
		NA	0	12	2	6	0	20
Sampling methods match	38.233 (p=0.002)	Reported	4 (33.33%)	35 (46.67%)	8 (23.53%)	12 (57.14%)	10 (83.33%)	69 (44.81%)
		Not reported	8 (66.67%)	40 (53.33%)	26 (76.47%)	9 (42.86%)	2 (16.67%)	85 (55.19%)
		NA	5	4	0	3	1	13

NAs are not included in percentages, to enhance readability and comparability with statistical testing.

Table 3 Treatment of data equivalence issues by time period

	t-test (sig.)	*Categories*	*1995–1999*	*2000–2005*	*Total*
			(n=60)	*(n=107)*	*(n=167)*
Construct equivalence					
Functional equivalence	0.733 (p=0.465)	Reported	19 (32.2%)	38 (38.0%)	57 (35.8%)
		Not reported	40 (67.8%)	62 (62.0%)	102 (64.2%)
		NA	1	7	8
Conceptual equivalence	0.321 (p=0.748)	Reported	23 (39.0%)	42 (41.6%)	65 (40.6%)
		Not reported	36 (61.0%)	59 (58.4%)	95 (59.4%)
		NA	1	6	7
Categorical equivalence	−0.362 (p=0.718)	Reported	12 (20.3%)	18 (18.0%)	30 (18.9%)
		Not reported	47 (79.7%)	82 (82.0%)	129 (81.1%)
		NA	1	7	8
Reliability	0.987 (p=0.325)	Reported	30 (50.0%)	62 (57.9%)	92 (55.1%)
		Not reported	30 (50.0%)	45 (42.1%)	75 (44.9%)
Unidimensionality	0.328 (p=0.743)	Reported	22 (36.7%)	42 (39.3%)	64 (38.3%)
		Not reported	38 (63.3%)	65 (60.7%)	103 (61.7%)
Construct validity	0.795 (p=0.428)	Reported	22 (36.7%)	46 (43.0%)	68 (40.7%)
		Not reported	38 (63.3%)	61 (57.0%)	99 (59.3%)
Measurement equivalence					
Calibration equivalence	−1.121 (p=0.264)	Reported	11 (20.0%)	12 (13.0%)	23 (15.6%)
		Not reported	44 (80.0%)	80 (87.0%)	124 (84.4%)
		NA	5	15	20
Translation equivalence	−0.228 (p=0.820)	Reported	28 (50.9%)	48 (49.0%)	76 (49.7%)
		Not reported	27 (49.1%)	50 (51.0%)	77 (50.3%)
		NA	5	9	14
Metric equivalence	−0.731 (p=0.466)	Reported	17 (28.30%)	24 (22.40%)	41 (24.60%)
		Not reported	41 (68.30%)	76 (71.00%)	117 (70.10%)
		NA	2	7	9
Scoring consistency	−0.731 (p=0.466)	Reported	14 (24.1%)	19 (19.2%)	33 (21.0%)
		Not reported	44 (75.9%)	80 (80.8%)	124 (79.0%)
		NA	2	8	10
Scalar equivalence	−0.121 (p=0.904)	Reported	11 (19.0%)	18 (18.2%)	29 (18.5%)
		Not reported	47 (81.0%)	81 (81.8%)	128 (81.5%)
		NA	2	8	10
Data collection equivalence					
Sampling frame comparability	−0.510 (p=0.611)	Reported	15 (28.3%)	26 (24.5%)	41 (25.8%)
		Not reported	38 (71.7%)	80 (75.5%)	118 (74.2%)
		NA	7	1	8
Data collection procedure	−0.225 (p=0.822)	Reported	30 (53.6%)	47 (51.6%)	77 (52.4%)
		Not reported	26 (46.4%)	44 (48.4%)	70 (47.6%)
		NA	4	4	8
Sampling methods match	1.185 (p=0.238)	Reported	22 (38.6%)	47 (48.5%)	69 (44.8%)
		Not reported	35 (61.4%)	50 (51.5%)	85 (55.2%)
		NA	3	10	13

NAs are not included in percentages, to enhance readability and comparability with statistical testing.

(Boyacigiller & Adler, 1991). During the original construct validation process, western researchers attached meanings from the respondents' scores to particular research instruments based upon their specific value systems as well as any potential actions that might result from their interpretations (Messick, 1995). At the country level, this implies that the more culturally distant the country is from

where original construct validation takes place, the theoretical underpinnings and empirical evidence have a lower probability of being equivalent. For example, as Guatemala is in an opposite quadrant for all of Hofstede's dimensions (e.g., Hofstede, 1980) *vis-à-vis* the US, one should assume that perceptual constructs are *not* equivalent between the two cultures until proven otherwise. Unfortunately, the results indicate that more than half the studies did not report meeting this standard of evidence. As a result, aside from creating confusion in the literature with divergent findings, cross-cultural IB research knowledge cannot effectively accumulate for the advancement of our discipline.

To overcome this limitation in the literature it is important that researchers first determine whether the phenomena under investigation actually exist and are interpreted similarly across cultures or in the countries studied (i.e., functional, conceptual, and category equivalence). In order to establish construct equivalence researchers should draw not only from the extant domestic literature when building the conceptualization of constructs, but also, to the extent possible, from the country-specific literature when developing conceptualizations. In this manner, a greater understanding of the emic and etic aspects of the constructs can be gained. Further, qualitative research (e.g., interviews, focus groups, pre-tests, and pilot studies) should be conducted to identify cultural differences attached to the meaning of the construct in each country under investigation (Kumar, 2000; Schwarz, 2003). Ultimately, researchers should work to ensure that the concept could be measured using similar questions in every country. If not, concepts should be operationalized using emic, or culturally specific, measures to represent the theoretical domain of the construct more fully (Bensaou, Coyne, & Venkatraman, 1999; Mullen, 1995).

In addition, the steps taken by researchers to establish functional, conceptual, and category equivalence should be reported in the articles developed. In an eight-country study, Harpaz, Honig and Coetsier (2002) provide a template of how to use qualitative and quantitative techniques effectively to reduce the risk of construct and measurement inequality between nations. Specifically, these scholars use an iterative approach to scale development in each country, where the items are initially derived from construct conceptualizations in each country and then are iteratively exposed to statistical testing, resulting in similarity of construct measurement across countries. Further, Bensaou et al. (1999) propose a conceptual and analytical framework for assessing measurement equivalence. They collected data in the automobile industries in the US and Japan. In their study, functional equivalence was satisfied because their focus was on interorganizational relationships, which have the same meaning in both countries with similar economic philosophies (i.e., capitalist). For conceptual equivalence, they conducted exploratory fieldwork in both countries in the local language and by the same researcher to investigate the way managers interpreted key concepts. Finally, category equivalence was established because the managers in both countries used the same language and categories to discuss key concepts of strategy and tactics within the industry. Once functional, conceptual, and category equivalence are established, data collection can proceed.

Post-data collection assessment procedures include assessments for unidimensionality, reliability and construct validity. Although various methods are available to IB researchers, the current application and reporting do not meet acceptable standards for the field to advance. In particular, the establishment of unidimensionality necessitates the establishment of convergent and discriminant validity, as unidimensionality means determining whether a set of indicators reflect one underlying factor. Convergent validity may be evaluated, at the construct level, by the average variance extracted of the construct explained by the indicators (AVE) and, at the item level, by examining the item-to-total and inter-item correlations and the factor loadings of the indicators. Tests of discriminant validity offer evidence (absence) of items cross-loading onto conceptually similar constructs. Assessment of discriminant validity, at both the construct and item levels, may be conducted by using a Lagrangian multiplier test (Anderson & Gerbing, 1988). Other techniques at the construct level include analysis of pairwise factor correlations, and examination of their confidence intervals (Gerbing & Anderson, 1992). Further, the most comprehensive cross-cultural assessment procedure for construct equivalence is a multi-group CFA (i.e., factorial similarity that pertains to scale items loading on the invariant factors in cross-cultural samples). More specifically, the choice of tools for establishing construct equivalence depends upon the sample, data and study characteristics, and the researcher's expertise.

MEASUREMENT EQUIVALENCE

Background

Measurement equivalence addresses the comparability of the operationalization of the constructs, that is, the wording, scaling, and scoring of measures across different populations (Mullen, 1995). Without the establishment of measurement equivalence the validity of findings is called into question (Horn, 1991; Steenkamp & Baumgartner, 1998). Measurement equivalence encompasses three critical components: calibration, translation, and metric equivalence (Craig & Douglas, 2000; Sekaran, 1983; Steenkamp & Baumgartner, 1998).

Prior to data collection, researchers should focus on the establishment of calibration equivalence (i.e., ensuring units of measure are converted correctly between cultures) and translation equivalence (i.e., ensuring questionnaire items are translated appropriately so that items tap into the same latent constructs in different populations) (Mullen, 1995). Calibration equivalence reflects equality between physical and perceptual measures across cultures, while translation equivalence reflects the conveyance of identical meaning from culture to culture. In the cross-cultural IB literature, back-translation has historically been the most commonly used method for the establishment of calibration and translation equivalence (Mullen, 1995), where back-translation procedures provide researchers with a language check and, more importantly, the compatibility of concepts between the national cultures can be assessed during the translation process (Sekaran, 1983). However, researchers should use caution in focusing on semantics, as literal translations of the measurements can become stilted and lacking in the naturalness required such that respondents understand the concept (van de Vijver & Leung, 1997). When the translations are not similar, or are incomparable because of cultural differences, concepts should be operationalized utilizing emic, or culturally specific, measures to represent the theoretical domain of the construct fully and equally across the cultures (Mintu et al., 1994).

Further, several measures are available to IB researchers to assess metric equivalence: for example, item-level checks ensure invariance and consistency of the subject responses to the measurement scales, post data collection. Metric equivalence has two important facets: consistency of scoring, and equality of responses (i.e., scalar equivalence) (Craig & Douglas, 2000). Inconsistency in scoring may arise from a lack of familiarity with scaling and scoring formats. For example, "American researchers tend to use 5 to 7 point Likert scales to measure perceptions, whereas researchers from Europe tend to use 20 point scales" (Kumar, 2000: 193). Hence variance in scoring entails a threat to reliability, since it may attenuate (or accentuate) parameter estimates and statistical tests. In addition, tests of scalar equivalence (i.e., equality of responses – mean equivalence) attempt to distinguish between responses to items due to "actual" cultural differences. Lack of scalar equivalence may originate from respondent bias due to cultural factors, and may add to systematic measurement error. Hence inequality of responses entails a threat to the validity of research results. For example, some cultures in Latin America are known to provide "extreme" responses, whereas Asian cultures tend to favor more neutral responses (Steenkamp & Baumgartner, 1998). As demonstrated by Wong, Rindfleisch, and Burroughs (2003), some popular American instrument design practices cause problems when used in certain cultures that differ greatly from the US in terms of values, customs, and language. In particular, reverse coding of certain survey items (a widely espoused practice in the US) leads to low reliabilities when sampling East Asian respondents, possible because of religious belief system differences underlying eastern *vs* western cultures (cf. Wong et al., 2003). Thus understanding, and planning for, the idiosyncrasies of each research context is critical to achieving metric equivalence.

Findings

Our analysis captured whether measurement equivalence was reported (as depicted in Tables 2 and 3). For each article, we examined the aspects of measurement equivalence described previously (i.e., calibration equivalence, translation equivalence, metric equivalence (inclusive of scoring consistency and scalar equivalence)), and coded whether researchers reported appropriate procedures. Our findings indicate that very few articles (16%) reported estimates of calibration equivalence, and although there were no statistically significant differences across groups (F=1.403, p=0.236), *AMJ*, *SMJ*, and *JIBS* were the leading journals in reporting treatment of these issues. As shown in Table 3, attention to calibration equivalence decreased slightly across the two time periods (20% in 1995–1999; 13% in 2000–2005 (t=−1.121, p=0.264)). The findings related to translation procedure reporting is more encouraging, with half

of the studies reporting addressing of this issue. However, significant differences were found across journals in relation to reporting of translation equivalence (F=3.638, p=0.007), with *AMJ* being significantly different from *JWB*. Unfortunately, only 26% of the studies adequately addressed metric equivalence, with *AMJ*, *SMJ*, and *JIBS* being the leading journals in reporting treatment of this issue (with significant differences across journals (F=3.220, p=0.014), with *JIBS* being significantly different from *JWB* (p=0.034)). The percentage of articles that conducted post-data collection assessment for metric equivalence decreased slightly over time (28% in 1995–1999; 22% in 2000–2005 (t=−0.731, p=0.264)). Further, in terms of the two specific elements of metric equivalence, 33 studies (21%) addressed scoring consistency and 29 (19%) assessed scalar equivalence. As shown in Table 2, *JIBS*, *AMJ*, and *SMJ* had above-average performance in relation to these two dimensions, with no statistical differences across journals in relation to scoring consistency (F=2.065, p=0.088) but differences across journals in relation to scalar equivalence (F=3.257, p=0.014), with a significant difference between *JWB* and *JIBS* (p=0.008). As shown in Table 3, the overall lack of attention to metric invariance was consistently low across the two time periods.

Implications

Despite the importance of measurement equivalence, and the heightened attention paid to it in the literature (e.g., Craig & Douglas, 2000; Mullen, 1995; Myers, Calantone, Page, & Taylor, 2000; Steenkamp & Baumgartner, 1998), the results indicate that all three aspects of measurement equivalence are rarely fully reported or established. In part, this may reflect difficulty in the implementation of the procedures to establish calibration, translation, or metric equivalence. However, as researchers note, cross-cultural comparisons are not meaningful if the numbers on the response scales or the items that the respondents are responding to have different meanings across cultures (Kumar, 2000; Mullen, 1995).

To begin, pre-data collection researchers must work toward the establishment of calibration and translation equivalence. Calibration equivalence can be established by independently checking conversions of measurement instrument items (Mullen, 1995). This necessitates that researchers correctly identify and independently verify the conversion of measures contained in items, to ensure their comparability. As noted previously, calibration equivalence is closely tied to translation equivalence to ensure comparability of measures across cultures. To minimize errors in interpretation it is important that researchers apply and report translation procedures. These include back-translation, translation by committee, and testing for form and meaning equivalence (cf. Brislin, Lonner, & Thorndike, 1973; Mullen, 1995; Sekaran, 1983; Sperber, Devellis, & Boehlecke, 1994). For example, in a three-country study, Robertson, Al-Khatib, Al-Habib, & Lanoue (2001) utilize back-translation (English to Arabic), and adapt items in each country for local idiom. If researchers can demonstrate not only that the measures used are calibrated consistently across groups, but also that the meanings taken from the items are equivalent, the results of a study have greater validity.

Post data collection, metric equivalence (inclusive of scoring consistency and scalar equivalence) can be assessed in a number of ways, including multi-group structural equation models (SEM), profile analysis, optimal scaling, and regression analysis, and by comparing the standard deviations and means of the subjects' responses over a large number of items across cultures (Bollen, 1989; Mullen, 1995; Myers et al., 2000; Salzberger, Sinkovics, & Schlegelmilch, 2001). Scoring consistency can be checked by comparing reliabilities between groups, or by examining factor loadings and measurement error variances. Since measurement error is common in cross-cultural research, and attenuates correlations, structural relationships must be adjusted for variations due to unequal reliabilities across cultures (Singh, 1995). To establish scalar equivalence, pooled analysis (i.e., deculturing data by standardizing the responses to each observable variable within each sample separately, and removing scaling factors from the measurements) and adjustment factors for differences in reliability are typically used (Davis et al., 1981; Durvasula, Andrews, Lysonski, & Netemeyer, 1993; Singh, 1995). For example, the extant literature suggests alternating least-squares optimal scaling, which is a general extension of principal component analysis for use with non-metric or mixed metric data (Mullen, 1995; Myers et al., 2000; Salzberger et al., 2001), or using multiple respondents to assess consistency of results (e.g., Calantone, Schmidt, & Song, 1996; Davis et al., 1981). However, the most prevalent approach to address scalar equivalence has become hierarchical diagnoses of invariance of measurement using

multiple group structural equation modeling (e.g., Mullen, 1995; Myers et al., 2000; Steenkamp & Baumgartner, 1998, 2001). For example, Steenkamp and Baumgartner (1998) provide detailed procedures for testing full and partial measurement invariance that offer a more comprehensive diagnosis than other techniques and simultaneous test of samples. The test begins with an assessment of configural invariance, which asks whether the same simple pattern of factor loadings is obtained in both samples. The second step comprises an examination of factor covariance invariance and nomological validity by constraining the correlations among the factors (e.g., $\Phi_{jk}^1=\Phi_{jk}^2=\ldots=\Phi_{jk}^G$) to be equal. Next, metric invariance tests the equivalence of metrics and scale intervals between countries by constraining factor loadings (e.g., $\Lambda^1=\Lambda^2=\ldots=\Lambda^G$) to be equal across countries. Fourth, both the factor covariances (e.g., $\Phi_{jj}^1=\Phi_{jj}^2=\ldots=\Phi_{jj}^G$) and factor loadings are checked to be equal, to examine whether factor structure is consistent across countries (some also refer to this as translation equivalence). Finally, the invariance of the measurement error variances is assessed by additionally constraining the error variances to be equal across the groups (e.g., $\Theta^1=\Theta^2=\ldots=\Theta^G$) (Steenkamp & Baumgartner, 2001). This micro-level final step refers to item/scalar equivalence, which is generally overlooked in the cross-cultural IB literature.

DATA COLLECTION EQUIVALENCE

Background

Data collection equivalence refers to whether the sources of data, the methods of eliciting data and the resulting samples are comparable across cultures, and can be viewed in relation to three elements: sampling frame comparability, data collection procedure, and sample comparability. Without the establishment of data collection equivalence the validity of findings is called into question, as one cannot eliminate the alternative explanation that differences in sample frame or in data collection methods of samples account for differences in results across cultures.

Sampling frame comparability refers to whether the samples drawn from different cultures parallel each other, and can be established pre-data collection. Sampling frame inconsistencies lead to unequal sampling errors across countries (Kumar, 2000), reduce construct validity, and threaten the accuracy of findings. It therefore becomes imperative that researchers carefully scrutinize the sample frames across countries from which they will draw their samples to minimize threats to the data. Further, cross-cultural IB researchers, while attempting to establish sampling frame comparability, also work to maximize non-sample-frame differences to investigate the specific phenomena of interest (e.g., Cavusgil & Das, 1997; Peterson, 2001). Sivakumar & Nakata (2001) offer an empirical technique to strengthen cross-cultural sample designs "by maximizing differences (between countries) on focal variable(s), while minimizing and/or controlling differences on non-focal variables" (p. 565).

Data collection procedure equivalence involves administration equivalence (e.g., telephone interviews, surveys, etc.), coverage comparability (i.e., match of the degree of generalizability) and lapse of time between data collection in different countries. The establishment of comparable data collection procedures minimizes threats to validity. While maintaining equivalence in data collection procedures would appear straightforward, differences across countries in regulation (e.g., whether or not telephone interviews are allowed), cultural norms (e.g., cultural differences in responsiveness to telephone, in-person, and mail survey administration), mail systems (e.g., reliability and timeliness of mail delivery), etc., often lead to differences in data collection procedures.

Sampling method match encompasses the establishment of equivalence among sampling method techniques (i.e., probability *vs* non-probability) and the match between the representativeness of the data collected in different cultures (Salzberger et al., 2001). One of the most challenging steps in cross-cultural studies is the selection of a representative sample, owing to difficulties in determining which subjects embody the central tendencies of the culture (Sekaran, 1983). Random selection from a representative sample ensures that uncontrolled, systematic errors do not bias estimators. Researchers often resort to selecting matched samples in the countries of investigation by adjusting their sampling techniques in different cultures.

Findings

As indicated in Table 2, in terms of sample frame comparability the results indicate that only 41 (25%) of the studies established this element of data equivalence, with *JIBS* and *SMJ* being above average (however, there are no statistically significant differences in the reporting of sample frame equivalence across journals (F=1.916, p=0.110)).

Table 4 A checklist for establishing data equivalence

	Diagnostic method	*Benefit*	*Concern*
Construct equivalence			
Functional equivalence	Literature review and use of validated survey instruments	Use of existing scales	Verify if items apply to context
	Qualitative fieldwork: interviews, focus groups, pretests, pilot groups	Establish reliable and valid measures indigenous to culture	Expensive and time consuming
Conceptual equivalence	Literature review and use of validated survey instruments	Use of existing scales	Verify if items apply to context
	Qualitative fieldwork: interviews, focus groups, pretests, pilot groups	Establish reliable and valid measures indigenous to culture	Expensive and time consuming
Category equivalence	Literature review and use of validated survey instruments	Use of existing scales	Verify if items apply to context
	Qualitative fieldwork: interviews, focus groups, pretests, pilot groups	Establish reliable and valid measures indigenous to culture	Expensive and time consuming
Unidimensionality	Exploratory factor analysis and confirmatory factor analysis: factor loadings and error variances are identical for each scale item		
Reliability	Cronbach's alpha Kuder – Richarson coefficient	Confirm whether a measurement instrument in one language and setting has the same internal consistency properties	Not appropriate for formative measures
Construct validity (convergent validity)	Average variance extracted	Construct level	
	Item-total and inter-item correlations	Item level	
	Factor loading in factor analysis	Item level	
Construct validity (discriminant validity)	Lagrangian multiplier test	Can establish validity at item and construct level	
	Pairwise factor correlations and examining confidence intervals	Construct level	
	Multi-group CFA	Most comprehensive approach	Requires large sample size
Measurement equivalence			
Calibration equivalence	Independently check conversions of measurement units	Ensure measurement units, standards, and procedures for objective quantitative data are comparable	May require exclusion of secondary data sources; not transparent on definition of units
Translation equivalence	Back-translation	Widely applied to check for translation accuracy	Focuses on semantics rather than connotations, naturalness, and comprehensibility
	Translation by committee	The cooperative effort improves quality when members have varying areas of expertise	Researcher may need additional evidence of quality of translation
	Statistical testing for form and meaning equivalence: item-total correlations, ANOVA	Based on empirical data to assess linguistic equivalence	Application of statistical test depends on form of model equation (linear/non-linear) and sampling distribution
Metric equivalence (scalar)	Multiple group CFA tests for: configural invariance, factor covariance invariance, factor loadings across groups, equality of measurement error variances	Offers comprehensive diagnosis of measurement equivalence	Requires multiple measures of constructs; sample sizes may be larger; missing data challenges

Table 4 *Continued*

	Diagnostic method	*Benefit*	*Concern*
	Multiple methods of measurement		Time consuming and expensive
	Optimal scaling (alternating least-squares optimal scaling)	Examines the comparability of measures from different populations	Interpretation of results depends on the judgment and experience of researchers
		Applies to qualitative, metric, or mixed metric data.	
	Profile analysis	Gives researchers insights into response set bias	Does not indicate whether the differences in means between groups are caused by real differences in the variables
Metric equivalence (scoring consistency)	Compare reliabilities among cultures	Can use small sample sizes (for pretests)	
	Multiple-group CFA test for equality of measurement error variances	Most comprehensive assessment	Requires multiple measures of constructs and large sample size
Data collection equivalence			
Sampling frame comparability	Systematic selection of number and which cultures	Theory driven; can avoid bias	Greater cost and risk of equivalence issues
	Similar selection of sample among cultures	Likelihood of parallel respondents to reduce bias	Availability of comparable media for identifying sample may restrict some studies
Data collection procedure	Non-response bias test		
	Similar data collection procedures among cultures	Similar approaches; avoids bias in how data collected	Variations can be justified and controls in place to address method bias
	Time of data collection in countries	Simultaneous collection of data reduces alternative explanations	Coordination difficult
Sample comparability (sampling method match)	Matching samples	Best for between-country comparability	Observed similarities/differences cannot be generalized to whole country, limited to specific group involved
	Probability-sampling techniques	Best for descriptive IB research seeking within-country representativeness	Variation in key demographic characteristics may make it difficult to interpret
	Statistical control: entering sociodemographic variables as covariates – analysis of covariance and multiple regression	Best for between-country comparability	

The results pertaining to time periods indicate that the number of studies reporting sample frame comparability fell slightly from 28% to 24% (t=−0.510, p=0.611). Seventy-seven studies (52%) were identified as reporting the establishment of equivalence of data collection procedures. Significant differences were not found across journals in relation to reporting of data collection procedure equivalence (F=2.403, p=0.053). The percentage of articles reporting data collection procedure equivalence decreased slightly over time (54% in 1995–1999; 52% in 2000–2005 (t=−0.225, p=0.822)).

Finally, in 69 studies (45%), the sampling frame techniques matched across cultures, with significant differences across journals (F=38.233, p=0.002; with *JWB* being different from *SMJ* (p=0.002)). *JIBS*, *MIR*, and *SMJ* performed above average on this dimension, as shown in Table 2. The percentage of studies that matched sampling frames increased from 39% during 1995 to 1999 to 48% during 2000 to 2005 (t=1.185, p=0.238).

Implications

The findings of this study in relation to all three elements of data collection equivalence are concerning as, without data collection equivalence, sampling or procedural differences cannot be eliminated as an explanation for observed differences. Further, invariance in sample frames, data collection procedures, and sampling methods is a crucial aspect of cross-cultural studies, because they impact on construct and measurement equivalence, degree of reliability, and validity of the data analysis of the findings. For example, though Gibson (1995) utilized a convenience sample, statistical controls were used and metric invariance was tested, thus providing greater validity for the findings of the study. For cross-cultural IB research to advance, data collection equivalence limitations need to be overcome.

First, we contend that researchers should enlist parallel respondents for each country, ensure matches among sampling frame techniques and procedures, and allow minimum lapses of time between data collection in the different cultures. To achieve this, we also suggest that researchers use the approach suggested by Reynolds et al. (2003). Reynolds et al. (2003) developed a typology of international research that provides implications for balancing within-country representativeness and between-country comparability. Their framework consists of four types of research (i.e., descriptive, comparative, contextual, and theoretical), based on research questions and sampling choices. Because descriptive international research is focused on behavior and environment in a series of independent single countries, within-country representativeness is required. Comparative international research, on the other hand, is concerned with comparing behavior in two or more countries or cultural contexts to scrutinize similarities and/or differences. Thus, this type of research entails between-country comparability and matching samples (or control for sample idiosyncrasies). For example, Peterson et al's (1995) study of role stress and conflict among managers recognized discrepancies in the data, and used demographics and organizational characteristics to adjust scores to ensure the comparability of the samples. Contextual international research is interested in phenomena that cross national boundaries, and thus benefits from representativeness of the specific cross-national group of interest. Finally, theoretical international research seeks to "examine the extent to which theories, models and constructs developed in one country are valid and applicable in other countries and cultural contexts" (Craig & Douglas, 2000: 163). In this type of research, comparability is important in order to be able to determine the cross-national stability of the model. In an illustration provided by Reynolds et al. (2003), Balabanis, Diamantopoulos, Mueller, and Melewar (2001), using a theory-testing approach, use comparable samples from the Czech Republic and Turkey to test their model, despite the non-representativeness of the national populations. In addition, we argue that researchers should try to achieve sufficient variance based on the goals of the study via appropriate levels of heterogeneity, and by using as large a sample size as possible (Peterson, 2001). Lack of sufficient variance may lead to misleading results (i.e., fit indices, loadings, and effect sizes) and violate the assumptions of multivariate analysis techniques. Finally, if sample matching is not possible, or is sacrificed for the sake of high response rates across countries, statistical controls should be used (cf. Cavusgil & Das, 1997).

To avoid bias in data collection techniques, van de Vijver & Leung (1997) provide suggestions to reduce and test for problems in data collection procedures. For personal interviews, these include adequate training of interviewers to recognize potential biases, random assignment of interviewers, and including interviewer's characteristics such as age, gender, and communication quality as covariates. Lenartowicz and Johnson (2002) provide a good example of reducing interviewer responses bias in a 12-country study in Latin America by recruiting local interviewers, conducting extensive training, supervising initial interviews, limiting the role of the interviewer to collection of data, and conducting random follow-up checks on 10% of the sample.

CONCLUSION

The purpose of this study was to examine the extent to which current cross-cultural IB research

deals effectively with data equivalence issues (cf. Craig & Douglas, 2000; Mullen, 1995; Sekaran, 1983; Steenkamp & Baumgartner, 1998). Overall, we found that the standards called for in past methodological discussions have not been met. Although it would be reasonable to expect improvement over time *vis-à-vis* data equivalence, our findings revealed no statistical differences in the treatment of data equivalence issues between the 1995–1999 and 2000–2005 periods. Based on our results, we conclude that the cross-cultural IB literature has placed insufficient emphasis on data equivalence, and that greater attention needs to be paid to such issues for the field to advance. The stakes are quite high, for without evidence of data equivalence, the confidence that readers can have in findings is substantially reduced. In aggregate, the effect is one of impeding progress in the cross-cultural area in particular, and in IB research overall. To facilitate improved practice, Table 4 offers researchers and gatekeepers (i.e., editors and reviewers) a checklist for enhancing data equivalence in cross-cultural IB research. Our hope is that researchers will build on Table 4 such that a similar examination to ours conducted in the next decade would be able to report a higher level of established data equivalence in research than was reported here.

One clear source of optimism for the IB field is the relative performance of its flagship journal, *JIBS*. Although it is important to stress that the absolute levels found generally did not meet the field's standards, it is worth noting that studies published in *JIBS* exhibited above-average treatment of data equivalence issues in 12 of the 15 issues, and was at or within 3% of the average in the remaining issues. However, even with these results, the overall low levels of reported data equivalence suggest an opportunity to greatly enhance the rigor, and thereby the quality, of cross-cultural research in the field of IB.

ACKNOWLEDGEMENTS

We appreciate the guidance provided by Editor-in-Chief Arie Lewin and the anonymous reviewers, as well as research funding from the Center for International Business Education and Research at Michigan State University (MSU-CIBER).

NOTES

[1]We have included 20 studies that were conducted in a single country, but surveyed companies or managers from different countries.

[2]The percentages do not total to 100% owing to rounding.

[3]The Tamhane test was used for *post hoc* comparisons. Further, it is important to note that there were significant differences in sample sizes across journals: thus the *post hoc* analysis should be viewed with caution.

REFERENCES

Anderson, J. C., & Gerbing, D. W. 1988. Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin*, 103(3): 411–423.

Balabanis, G., Diamantopoulos, A., Mueller, R. D., & Melewar, T. C. 2001. The impact of nationalism, patriotism and internationalism on consumer ethnocentric tendencies. *Journal of International Business Studies*, 32(1): 157–175.

Bensaou, M., Coyne, M., & Venkatraman, N. 1999. Testing metric equivalence in cross-national strategy research: An empirical test across the United States and Japan. *Strategic Management Journal*, 20(7): 671–689.

Bollen, K. A. 1989. *Structural equations with latent variables*. New York: John Wiley & Sons.

Boyacigiller, N. A., & Adler, N. J. 1991. The parochial dinosaur: Organizational science in a global context. *Academy of Management Review*, 16(2): 262–290.

Brislin, R. W., Lonner, W. J., & Thorndike, R. M. 1973. *Cross-cultural research methods*. New York: John Wiley.

Calantone, R. J., Schmidt, J. B., & Song, X. M. 1996. Controllable factors of new product success: A cross-national comparison. *Marketing Science*, 15(4): 341–358.

Campbell, D. T., & Fiske, D. W. 1959. Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin*, 56(2): 81–105.

Cavusgil, S. T., & Das, A. 1997. Methodological issues in empirical cross-cultural research: A survey of the management literature and a framework. *Management International Review*, 37(1): 71–96.

Churchill Jr, G. A. 1979. A paradigm for developing better measures of marketing constructs. *Journal of Marketing Research*, 16(1): 64–73.

Combs, J. G., & Ketchen, D. J. 2003. Why do firms use franchising as an entrepreneurial strategy? A meta-analysis. *Journal of Management*, 29(3): 443–465.

Craig, C. S., & Douglas, S. P. 2000. *International market research*, 2nd edn. Chichester: John Wiley & Sons.

Cronbach, L. J. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3): 297–334.

Davis, H. L., Douglas, S. P., & Silk, A. J. 1981. Measure unreliability: A hidden threat to cross-national marketing research? *Journal of Marketing*, 45(2): 98–109.

DuBois, F. L., & Reeb, D. 2000. Ranking the international business journals. *Journal of International Business Studies*, 31(4):
689–704.

Durvasula, S., Andrews, J. C., Lysonski, S., & Netemeyer, R. G. 1993. Assessing the cross-national applicability of consumer behavior models: A model of attitude toward advertising in general. *Journal of Consumer Research*, 19(4): 626–636.

Fornell, C., & Larcker, D. F. 1981. Structural equation models with unobservable variables and measurement error:

Algebra and statistics. *Journal of Marketing Research*, 18(3): 382–388.
Gerbing, D. W., & Anderson, J. C. 1992. Monte Carlo evaluations of goodness-of-fit indices for structural equation models. *Sociological Methods and Research*, 21(2): 132–160.
Gibson, C. B. 1995. An investigation of gender differences in leadership across four countries. *Journal of International Business Studies*, 26(2): 255–279.
Harpaz, I., Honig, B., & Coetsier, P. 2002. A cross-cultural longitudinal analysis of the meaning of work and the socialization process of career starters. *Journal of World Business*, 37(4): 230–244.
Hofstede, G. 1980. *Culture's consequences: International differences in work-related values*. Newbury Park, CA: Sage.
Horn, J. L. 1991. Comments on issues in factorial invariance. In L. M. Collins and J. H. Horn (Eds) *Best methods for the analysis of change*: 114–125. Washington, DC: American Psychological Association.
Kumar, V. 2000. *International marketing research*. Upper Saddle River, NJ: Prentice-Hall.
Lenartowicz, T., & Johnson, J. P. 2002. Comparing managerial values in twelve Latin American countries: An exploratory study. *Management International Review*, 42(3): 279–307.
Messick, S. 1995. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9): 741–749.
Mintu, A. T., Calantone, R. J., & Gassenheimer, J. B. 1994. Towards improving cross-cultural research: Extending Churchill's research paradigm. *Journal of International Marketing*, 7(2): 5–23.
Mullen, M. R. 1995. Diagnosing measurement equivalence in cross-national research. *Journal of International Business Studies*, 26(3): 573–596.
Myers, M. B., Calantone, R. J., Page, T. J., & Taylor, C. R. 2000. Academic insights: An application of multiple-group causal models in assessing cross-cultural measurement equivalence. *Journal of International Marketing*, 8(4): 108–121.
Nunnally, J. C. 1978. *Psychometric theory*. New York: McGraw-Hill.
Peterson, R. A. 2001. On the use of college students in social science research: Insights from a second-order meta-analysis. *Journal of Consumer Research*, 28(3): 450–461.
Peterson, M. F., Smith, P. B., Akande, A., Ayestaran, S., et al. 1995. Role conflict, ambiguity, and overload: A 21-nation study. *Academy of Management Journal*, 38(2): 429–452.
Reynolds, N. L., Simintiras, A. C., & Diamantopoulos, A. 2003. Theoretical justification of sampling choices in international marketing research: Key issues and guidelines for researchers. *Journal of International Business Studies*, 34(1): 80–89.
Robertson, C., Al-Khatib, J., Al-Habib, M., & Lanoue, D. 2001. Beliefs about work in the Middle East and the convergence versus divergence of values. *Journal of World Business*, 36(3): 223–244.
Salzberger, T., Sinkovics, R. R., & Schlegelmilch, B. B. 2001. Data equivalence in international research: A comparison of classical test theory and latent trait theory based approaches. *Australasian Marketing Journal*, 7(2): 23–38.
Schwarz, N. 2003. Self-reports in consumer research: The challenge of comparing cohorts and cultures. *Journal of Consumer Research*, 29(4): 588–594.
Sekaran, U. 1983. Methodological and theoretical issues and advancements in cross-cultural research. *Journal of International Business Studies*, 14(2): 61–73.
Singh, J. 1995. Measurement issues in cross-national research. *Journal of International Business Studies*, 26(3): 597–619.
Sivakumar, K., & Nakata, C. 2001. The stampede toward Hofstede's framework: Avoiding the sample design pit in cross-cultural research. *Journal of International Business Studies*, 32(3): 555–574.
Sperber, A. D., Devellis, R. F., & Boehlecke, B. 1994. Cross-cultural translation: Methodology and validation. *Journal of Cross-Cultural Psychology*, 25(4): 501–524.
Steenkamp, J.-B. E., & Baumgartner, H. 1998. Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, 25(1): 78–90.
Steenkamp, J.-B. E., & Baumgartner, H. 2001. Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research*, 38(2): 143–156.
Tahai, A., & Meyer, M. H. 1999. A revealed preference study of management journals' direct influences. *Strategic Management Journal*, 20(3): 279–296.
van de Vijver, F., & Leung, K. 1997. *Methods and data analysis for cross-cultural research*. Thousand Oaks, CA: Sage Publications.
Wong, N., Rindfleisch, A., & Burroughs, J. E. 2003. Do reverse-worded items confound measures in cross-cultural consumer research? The case of the material values scale. *Journal of Consumer Research*, 30(1): 72–91.
Wright, R. W. 1970. Trends in international business research. *Journal of International Business Studies*, 1(1): 109–123.

APPENDIX

Coding Form

Construct equivalence: Are we studying the same phenomena in countries X, Y, and Z?

Functional equivalence: Checked whether a given concept or behavior serves the same function from country to country (including literature review).

Conceptual equivalence: Checked whether the same concepts/behaviors occur in different countries – the way in which they are expressed is similar.

Category equivalence: Checked same product attributes/characteristics considered.

Post data collection: Checked for tests for unidimensionaliy, reliability, convergent validity, discriminant validity

Any specifically national or cultural constructs?

Measurement equivalence: Are the phenomena in countries X, Y, and Z measured in the same way?

Calibration equivalence: For example, monetary units, measures of weights, distance and volume and perceptual cues; compared factor loadings (λ's) (via multigroup SEM); checked for the comparability of standards and units.

Translation equivalence: Checked whether the concept can be measured by using the same or similar questions in every country.

Method of translation/back-translation reported.

Metric equivalence: Checked for *scoring consistency*: compared reliabilities or compared measurement error variances (δ's).

Checked for scaling equivalence: Multimethod of measurement, profile analysis, optimal scaling or compared measurement error variances (δ's) and factor loadings (λ's) (via multigroup SEM).
Decentered or adapted scoring/scaling for that country?
Any other cultural biases (e.g., exaggerated or mean responses) accounted for?
Data collection equivalence: Are the data collection procedures in countries X, Y, and Z the same?
Criterion for country/culture selection (convenience, theoretical justification); sufficient variance between countries/cultures.
Sample size for each country/culture studied.
Relevant or same respondent for each country (manager, decision-maker, executive, etc.).
Sampling frame techniques match between countries?
Sampling frame comparability.
Coverage comparability.
Countries where the survey was developed.
Sampling procedure equivalence (telephone interviews, surveys, etc.) (do procedures match?).
Any procedure for non-sampling/non-response bias?
Sampling method in each country

ABOUT THE AUTHORS

G Tomas M Hult (PhD, University of Memphis) is Professor of Marketing and International Business and Director of the Center for International Business Education and Research (MSU-CIBER) in the Eli Broad Graduate School of Management at Michigan State University. He is Executive Director of the Academy of International Business, and has held roles as Associate Editor-in-Chief and Department Editor of the *Journal of International Business Studies*. He is a native of Sweden. E-mail: hult@msu.edu.

David J Ketchen Jr (PhD, Penn State University) is the Lowder Eminent Scholar and Professor of Management at Auburn University. He is a former Department Editor of the *Journal of International Business Studies* and currently serves as an associate editor for the *Academy of Management Journal*. His research focuses on strategy, entrepreneurship, and research methods. He is a native of the US. E-mail: ketchda@auburn.edu.

David A Griffith (PhD, Kent State University) is Associate Professor of Marketing in the Eli Broad Graduate School of Management at Michigan State University. His research interests include international marketing strategy and the employment of firm resources for strategic marketing effectiveness. He has published in the *Journal of Marketing*, *Journal of International Business Studies*, *Journal of Operations Management*, etc. He is a native of the US. E-mail: griffith@bus.msu.edu.

Carol A Finnegan (PhD, Michigan State University) is Assistant Professor of Marketing at the University of Colorado at Colorado Springs. She received her PhD from Michigan State University and MBA from Santa Clara University. Her research interests include international retail strategy and channel relationships. She is a native of the US. E-mail: cfinnega@uccs.edu.

Tracy Gonzalez-Padron is a PhD candidate in Marketing and International Business in the Eli Broad Graduate School of Management at Michigan State University. Her research interests include marketing strategy, supply chain management, global marketing, and corporate social responsibility. She has published in *Industrial Marketing Management* and various conferences. She is a native of the US. E-mail: gonza297@msu.edu.

Nukhet Harmancioglu (PhD, Michigan State University) is Assistant Professor of Marketing at Bilkent University. Her research interests include strategic marketing, new product development, and international business. She has won research awards from both the Product Development Management Association and Academy of Marketing Science, and has published in the *Journal of Product and Innovation Management* and *R&D Management*. She is a native of Turkey. E-mail: nukheth@bilkent.edu.tr.

Ying Huang (PhD, Michigan State University) is Assistant Professor of Retailing and Consumer Sciences at the University of Arizona. Her research interests include retailer – supplier relationships, retail new product development, and retailers' international expansions. She has published in *International Business Review* and various conferences. She is a native of China. E-mail: huang2@email.arizona.edu.

Mehmet Berk Talay (PhD, Michigan State University) is Assistant Professor of Marketing, Department of Marketing, HEC Montréal. His research focuses on innovation, new product development, and the dynamics of competition and coevolution of firms and products. His research has appeared in such journals as *Journal of Product Innovation Management*, *Industrial Marketing Management*, and *Journal of Marketing Theory and Practice*. He is a native of Turkey. E-mail: berk.talay@hec.ca.

S Tamer Cavusgil (PhD, University of Wisconsin) is University Distinguished Faculty and The John W. Byington Endowed Chair in Global Marketing in the Eli Broad Graduate School of Management at Michigan State University. His areas of interest include international marketing strategy, early internationalization, and emerging markets. He served as the inaugural Editor of *Journal of International Marketing*, and edits *Advances in International Marketing*. He is a native of Turkey. E-mail: cavusgil@msu.edu.

Accepted by Arie Y Lewin, Editor-in-Chief, 21 August 2007. This paper has been with the authors for two revisions.