Organizational Research Methods
Volume 9 Number 3
July 2006 339-368

10.1177/1094428106287434
http://orm.sagepub.com
hosted at
http://online.sagepub.com

The Measurement Equivalence of Web-Based and Paper-and-Pencil Measures of Transformational Leadership

A Multinational Test

Michael S. Cole
University of St. Gallen
Arthur G. Bedeian
Louisiana State University
Hubert S. Feild
Auburn University

Multigroup confirmatory factor analysis was applied to the responses of 4,909 employees of a multinational organization with locations in 50 countries to examine the measurement equivalence of otherwise identical Web-based and paper-and-pencil versions of 20 items comprising the transformational leadership component of Bass and Avolio's Multifactor Leadership Questionnaire. The results supported configural, metric, scalar, measurement error, and relational equivalence across administration modes, indicating that the psychometric properties of the 20 items were similar whether administered as a paper-and-pencil or Web-based measure. Although caution is always advised when considering multiple modes of administration, the results suggest that there are minimal measurement differences for well-developed, psychometrically sound instruments applied using either a paper-and-pencil or an online format. Thus, the results open a methodological door for survey researchers wishing to (a) assess transformational leadership with a Web-based platform and (b) compare or combine responses collected with paper-and-pencil and Web-based applications.

Keywords: *transformational leadership; measurement invariance*

More than a decade has passed since Synodinos and Brennan's (1990) prediction that, beyond radically changing the nature of business research, interactive computerized software would become the preferred method for collecting and analyzing survey responses. The combined promise of reduced administration expenses and shortened collection-analysis-feedback cycles has undoubtedly encouraged an increasing number of organizations to adopt computer-based research techniques (Hogg, 2002; Thompson, Surface, Martin, & Sanders, 2003). Contrasting opinions regarding the efficacy of computerized (non-Web and Web-based) versus conventional paper-and-pencil surveys have been voiced for more than

Authors' Note: We thank Hettie A. Richardson for vetting an interim draft manuscript. Please address correspondence to Michael S. Cole, Institute for Leadership and Human Resource Management, University of St. Gallen, Dufourstrasse 40a, St. Gallen, Switzerland CH-9000; e-mail: michael.cole@unisg.ch.

30 years (Buchanan, 2002; Evan & Miller, 1969; Gosling, Vazire, Srivastava, & John, 2004). Our intent is not to revisit the debate surrounding use of computer-based surveys but to empirically address a broader issue that is a precondition for the meaningful comparison (or aggregation) of survey responses elicited using paper-and-pencil and Web-based measures. Moving beyond basic considerations associated with measurement reliability and validity, we explore the measurement of paper-and-pencil and Web-based versions of an established measure of transformational leadership. In doing so, we place transformational leadership within a nomological network composed of three theoretically related workplace constructs (viz., collective efficacy, work-group cohesiveness, and collective goal commitment). Questions associated with measurement equivalence (or, alternatively, measurement invariance) include whether items asked on paper-and-pencil and Web-based versions of a measure are conceptually equivalent, whether paper-and-pencil and Web-based operationalizations of underlying theoretical constructs yield equivalent associations, and whether responses collected using paper-and-pencil and Web-based measures are subject to the same forms of nonsystematic measurement error.

Previous Research

Whenever a survey measure is converted for use online, it is important for scientific inference, not to mention being an ethical responsibility, to have evidence of measurement equivalence (ethical principles 9.02b and 9.05, American Psychological Association, 2002; see also Naglieri et al., 2004). Buchanan, Johnson, and Goldberg (2005) recently remarked that "it is clear that one cannot simply mount an existing [measure] on the World Wide Web and assume that it will be exactly the same instrument" (p. 125). Moreover, lack of evidence of measurement invariance weakens conclusions because findings are open to alternative interpretations (Steenkamp & Baumgartner, 1998).

As was the case in the present study, most Web-based instruments are simple translations of survey measures developed using a more traditional paper-and-pencil format. Based on their review of the literature, Buchanan and colleagues concluded that although such Web-based surveys can be reliable and valid, there is ample empirical evidence to suggest that their psychometric properties cannot be taken for granted (e.g., Buchanan, 2002; Buchanan & Smith, 1999). Fouladi, McCarthy, and Moller (2002) reported, for example, that a slight format modification adversely affected the psychometric properties of a Web-based version of two established individual-difference measures. Likewise, in a study of the viability of a Web-based personality inventory based on the Big Five taxonomy, Buchanan et al. (2005) found some items loaded on the "wrong" factors or had unacceptable cross-factor loadings. They concluded that because factorial validities of multidimensional measures are prone to subtle changes when administered online, not all Web-based and paper-and-pencil versions of the same measure will be equivalent.

Other evidence suggests, however, that mode of administration may not adversely affect the comparability of responses to organizationally relevant, job attitude measures. Potosky and Bobko (1997), for instance, using a sample of undergraduate students ($n = 176$) as participants in a simulated personnel-selection process, reported cross-mode correlations

(respondents completed both paper-and-pencil and computer-based versions of each measure) of 1.0 when corrected for unreliability. A repeated-measures comparison of mean respondent scores also indicated there were no significant differences across administration modes. Stanton (1998) investigated the comparability of procedural and distributive justice items using two employee samples. Sample 1 (n = 50) completed a Web-based survey, whereas Sample 2 (n = 181) completed a paper-and-pencil instrument. He investigated both the measurement equivalence of the measures across modes and the independent samples' mean differences. Although use of a covariance analysis framework is generally not recommended for very small samples, Stanton found that the measures' factor structures, item loadings, correlations between factors, and latent factor variances were equivalent across administration modes. Moreover, he found that respondents who completed the paper-and-pencil survey reported higher levels of fairness.

Similarly, Donovan, Drasgow, and Probst (2000) used an item-response theory (IRT) technique to examine the measurement equivalence of two dimensions of the Job Descriptive Index (viz., supervisor and coworker satisfaction) across three employee samples. Two samples (n = 1,777) were administered a paper-and-pencil instrument, and the third (n = 509) was administered a computer-based version of the same instrument. Results were mixed with regard to levels of measurement equivalence across the paper-and-pencil and computer administrations. Donovan et al. acknowledged, however, that the university employees who received the computerized instrument were distinctly different from the blue-collar employees who completed the paper-and-pencil instrument, and therefore, sample characteristics were confounded with modes of administration. In light of this limitation, Donovan et al. recommended that researchers obtain multiple samples from the same organization to permit an unconfounded assessment of measurement equivalence.

Considered collectively, research on the measurement equivalence of responses collected using different administration modes has produced mixed findings and is limited in quantity. For these reasons, Simsek and Veiga (2001) and, more recently, Thompson et al. (2003) have concluded that research exploring measurement invariance across paper-and-pencil and Web-based surveys remains in its infancy. Moreover, as will be discussed, the existing research is methodologically suspect.

Study Purpose

Based on a review of the methodological challenges facing researchers, Stanton and Rogelberg (2001) resolved that research is still needed to fully understand how different methods for presenting research stimuli influence survey responses. Despite continued calls to address the issue of measurement invariance (Millsap & Kwok, 2004; Vandenberg, 2002), organizational researchers have seemingly assumed that paper-and-pencil and Web-based surveys exhibit adequate cross-mode equivalence. As others have noted, however, a failure to demonstrate measurement equivalence across administration modes not only casts doubt on their comparability but also may make their unambiguous interpretation impossible (Vandenberg & Lance, 2000). Therefore, results from earlier studies, whether as simple as comparing mean scores across administration modes or as complex as comparing or testing

measurement models based on data from paper-and-pencil or Web-based surveys, may be at best equivocal (Riordan & Vandenberg, 1994) and at worst erroneous (Cheung & Rensvold, 2002; Steenkamp & Baumgartner, 1998). Others have noted this concern and similarly argued that measurement equivalence of Web-based versions of paper-and-pencil surveys cannot be assumed but must be empirically demonstrated (Booth-Kewley, Edwards, & Rosenfeld, 1992; Buchanan & Smith, 1999; Cohen, Swerdlik, & Phillips, 1996).

The importance of measurement equivalence is underscored when it is realized that if, for example, a measure's factor structure and pattern of factor loadings were found to vary across paper-and pencil and Web-based applications (and, thus, perhaps be as much a function of a measure's administration mode as of its content), as noted above, there would be no scientific basis for drawing meaningful inferences or establishing the generalizability of competing theories. Thus, the general question of measurement invariance must be addressed when comparing or aggregating data collected using Web-based and paper-and-pencil surveys, when selecting one administration mode over another or when switching between modes. Moreover, in that measurement equivalence highlights the question of factorial validity, it is important to examine whether administration mode affects structural relations among theoretical constructs comprising a nomological network, or what Drasgow (1984) has dubbed "relational equivalence." In extending Drasgow's logic, if a measure provides equivalent relations across administration modes, "it is natural to ask whether [a measure's] scores have comparable relations with other important variables" (p. 134).

Beyond the preceding considerations, as noted by Ryan, Chan, Ployhart, and Slade (1999), there has been little research directed toward the measurement of job attitudes in multinational organizations. The purpose of the present investigation was therefore to examine measurement equivalence issues and to do so within the context of an international study involving employees representing 50 countries. A complete understanding of measurement issues associated with multinational research and further advancement of management as an academic discipline requires that the psychometric properties of instruments developed in one country be evaluated in other countries to establish that they are cross-nationally invariant. As Steenkamp and Baumgartner (1998) observed, cross-national differences in measurement properties may, for example, be due to actual differences among respondents from different countries on an underlying theoretical construct or "may be due to systematic biases in the way people from different countries respond to certain items" (p. 78).

Guided by the prevailing assumptions that transformational leadership will (a) continue to exist as a focal research topic, (b) be of growing interest to international researchers, and (c) be increasingly assessed using Web-based methods, we surmised that an overdue contribution to the measurement literature would be to empirically test if the most frequently used measure of transformational leadership, taken from Bass and Avolio's (2000) Multifactor Leadership Questionnaire (TL-MLQ), exhibited comparable psychometric properties across paper-and-pencil and Web-based surveys. Moreover, as an extension of previous measurement invariance research, we sought to examine whether administration mode affects whether TL-MLQ scores have equivalent relations with three theoretically related workplace constructs (viz., collective efficacy, work-group cohesiveness, and collective goal commitment). Evidence that (a) commonly accepted measures of theoretical

constructs are cross-nationally equivalent and (b) the structural relations among these constructs as elements comprising transformational leadership's nomological network are comparable across divisions of a multinational organization would support use of either paper-and-pencil or Web-based versions of the Bass and Avolio's transformational leadership measure in international research.

To accomplish our stated purpose, we evaluated the measurement equivalence of otherwise identical Web-based and paper-and-pencil versions of Bass and Avolio's (2000) transformational leadership measure as well as established measures of three work-related constructs (identified above) within a multigroup confirmatory factor analysis (MGCFA) framework. In doing so, we sought to establish their measurement equivalence within a hypothesized nomological network composed of transformational leadership (Bass & Avolio, 2000) and collective efficacy, work-group cohesiveness, and collective goal commitment (Zaccaro, Rittman, & Marks, 2001).

MGCFA

MGCFA is a powerful and versatile approach for testing measurement equivalence across applications (Cheung & Rensvold, 2002; Meade & Lautenschlager, 2004b; Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000). As such, it offers several advantages (Vandenberg, 2002). First, the basic CFA framework accounts for measurement error through the inclusion of error terms. Second, by varying constraints across a series of nested models, MGCFA permits a direct examination of measurement equivalence. Thus, by using MGCFA, we were able to specify constraints in an a priori manner, which allowed us to test for measurement equivalence across administration modes. Finally, MGCFA equivalence analyses, compared to their IRT counterpart, test across groups simultaneously and therefore are preferred when multiple groups are being compared (as was the case in several of our analyses) and sample sizes are small (Meade & Lautenschlager, 2004a).

An in-depth discussion of the recommended procedures for testing measurement equivalence is available in Vandenberg and Lance (2000) and Vandenberg (2002). In brief, when testing for measurement equivalence across applications (or, as in the present instance, administration modes), sets of parameters are constrained in a logically ordered, increasingly restrictive fashion. Little (1997) categorized these measurement equivalence tests into two types. Measurement-level tests assess (a) configural, (b) metric, (c) scalar, and (d) measurement-error equivalence across applications. Configural equivalence evaluates whether the conceptual frame of reference used by respondents exposed to different applications are comparable and is operationalized by testing for similarity in the pattern of factor loadings across applications. It is generally considered the weakest form of equivalence. Metric invariance assesses whether factor loadings for like items are equal across applications. Some level of metric invariance must be evident for subsequent tests of measurement equivalence to be interpretable. Scalar (or intercept) equivalence involves testing whether the vector of intercepts for like items is invariant across applications. Scalar equivalence suggests that like items have the same operational definition across applications (Cheung & Rensvold, 2002). Finally, measurement-error equivalence assesses whether, across applications, like items tap the same underlying factors with a similar degree of measurement error.

Latent-construct tests assess construct-level equivalence and involve testing for invariant construct variances/covariances, as well as between-application differences in latent means (for an in-depth discussion of mean and covariance structures analysis, see Ployhart & Oswald, 2004). Whereas some methodologists have suggested that latent-construct-level tests should be conducted only in the presence of complete support for configural, metric, and scalar equivalence (e.g., Cheung & Rensvold, 2002), others have advocated relaxing certain measurement-level constraints and testing latent means under conditions of partial scalar equivalence (for a discussion of partial invariance, see Byrne, Shavelson, & Muthén, 1989; Cheung & Rensvold, 1999; Millsap & Kwok, 2004; Steenkamp & Baumgartner, 1998).

To guide the present study, we followed the sequence of test procedures recommended by Vandenberg and Lance (2000; Vandenberg, 2002). Specifically, we compared the number of underlying factors and equality of the item-factor loadings for measures of transformational leadership, collective efficacy, work-group cohesiveness, and collective goal commitment administered in both traditional pencil-and-paper and Web-based versions. Where both configural and metric equivalence were supported across administration modes, we tested for scalar equivalence. Finding scalar equivalence across applications, we then tested for equality of residuals to establish measurement-error equivalence. With item uniqueness supported across applications, we then tested for relational equivalence by exploring the equality of the relevant latent-construct variances and covariances. Finally, latent-mean comparisons were conducted to test for complete measurement equivalence.

Method

Research Setting

Targeted respondents were employed by a company involved in the manufacture, renovation, and service of power generators and equipment. In total, 8,598 employees located in 50 countries were invited to participate in the present study. Given that the employees varied in their ability to comprehend English, all study measures were translated into 16 additional languages by a professional translation service. Linguists followed a blind, back-translation strategy. That is, one linguist translated each measure from English into a second language; then, another linguist back-translated this second language into English. The two linguists discussed any discrepancies and modified the measures accordingly.

The company was adamant in its desire that every employee receive an invitation to participate in the study. To ensure that all employees had equal study access, both Web-based and paper-and-pencil surveys incorporating the focal measures were developed. Every employee with a company e-mail address (n = 7,319) received an electronic message from the company chief executive officer (CEO) describing the study and providing a link to a portal (hosted by the first author's university) where a Web-based version of the survey could be completed in any of 17 languages. In addition, 1,279 employees who did not have a company e-mail address or computer access were sent an identical, paper-and-pencil survey using the company's internal mail service. In all, 4,909 employees chose to voluntarily

participate in the study by completing either a Web-based or paper-and-pencil survey, yielding a 57% response rate.

Participants

Web-based sample. In the online sample, 4,244 employees completed surveys for a 58% response rate. The sample consisted of 12% upper-level managers, 29% line managers and team leaders, and 59% nonmanagerial employees. The majority (75%) were between the ages of 30 and 55, and more than half (66%) had worked for the company for more than 5 years. Age ranges, organization tenure, job functions, job titles, and survey language percentages are shown in Table 1.

Paper-and-pencil sample. In the paper-and-pencil sample, 665 employees completed surveys for a 52% response rate. The paper-and-pencil sample included 4% upper-level managers, 8% line managers and team leaders, and 88% nonmanagerial employees. The majority (73%) were between the ages of 30 and 55, and more than half (74%) had worked for the company for more than 5 years. Age ranges, organization tenure, job functions, job titles, and survey language percentages are shown in Table 1.

Although the two samples varied somewhat with regard to demographic characteristics, such differences were anticipated because of the company's multinational structure and geographically dispersed workforce. As one would expect, a plurality of upper-level managers were located in the company's state-of-the-art home office, whereas a majority of its manufacturing employees were spread worldwide in facilities with varying electronic access. It is this difference in access that created the conditions necessary for the present study to be conducted. There is no theoretical reason, however, to believe that such differences would be systematically related to the focal variables of interest (described below) and thus affect relationships among study variables.

Measures

Transformational leadership. Respondents assessed their direct supervisor's leadership behavior using the 20 transformational leadership items of the MLQ-5x-short (Bass & Avolio, 2000), the most frequently used measure of transformational leadership (Antonakis, Avolio, & Sivasubramaniam, 2003). The MLQ-5x-short assesses four dimensions of transformational leadership: idealized influence (attributed and behavior), inspirational motivation, intellectual stimulation, and individualized consideration. Using a 5-point response format ranging from 1 = *not at all* to 5 = *frequently, if not always*, respondents judged how frequently their supervisors displayed specific leader behaviors. Coefficient alpha for the paper-and-pencil version was .96; alpha for the Web-based version was .96.

Collective efficacy. Six items from Riggs and Knight's (1994) collective-efficacy beliefs measure were used to substantiate respondents' individual perceptions of their work group's competence for successful action. Sample items include, "People in my work group

Table 1
Demographic Information for Web-Based and Paper-and-Pencil Samples

Variable	Web-Based Sample (n = 4,244)	Paper-and-Pencil Sample (n = 665)
Response rate (%)	58.0	52.0
Age in years (%)		
<25	2.2	3.2
25-30	10.7	11.1
31-35	15.8	9.2
36-40	15.6	17.7
41-45	14.0	15.6
46-50	15.0	18.5
51-55	14.1	12.2
56-60	7.3	4.7
>60	2.9	1.5
No answer	2.5	6.3
Organization tenure in years (%)		
<1	3.9	3.0
<5	28.5	16.5
5-10	19.6	22.3
11-15	13.1	20.2
16-20	7.0	8.3
21-25	9.4	6.0
26-30	7.8	7.8
31-35	5.5	6.6
>35	3.3	2.9
No answer	1.8	6.5
Job function (%)		
Field service	16.6	9.5
Workshops	4.8	51.6
Maintenance	8.5	9.6
Engineering	22.9	3.2
Sales	11.4	3.9
Parts and logistics	6.1	3.6
Support functions	22.3	5.1
Shared services	4.0	.6
Other	3.5	12.9
Job title (%)		
Sector/region/unit management	11.8	3.9
First line manager/team leader	29.4	7.5
Employee	58.9	88.6

(continued)

have above average ability" and "My work group is not able to perform as well as it should" (reverse scored). Respondents indicated their agreement with these items using a 5-point response format (1 = *strongly disagree,* 5 = *strongly agree*). The paper-and-pencil version alpha was .74; alpha for the Web-based version was .77.

Table 1 (continued)

Variable	Web-Based Sample (n = 4,244)	Paper-and-Pencil Sample (n = 665)
Survey language (%)		
Croatian	4.2	35.9
Czech	1.5	5.9
Danish	1.8	10.7
English	39.0	10.5
Estonian	0.2	0.2
Finnish	0.9	1.2
Flemish	—	1.8
French	13.4	4.8
German	15.6	3.2
Hungarian	2.3	—
Italian	3.7	5.6
Japanese	0.1	—
Polish	3.0	13.2
Portuguese	2.4	—
Romanian	2.8	—
Spanish	3.8	1.7
Swedish	5.2	5.4

Work-group cohesiveness. Six items adapted from Riordan and Weatherly (1999) were used to gauge employees' work-group cohesiveness. The items assess the extent to which a work group's members agree or disagree that they are willing to work together and are committed to completing assigned tasks. Sample items include, "In my work group, members know that they can depend on one another" and "In my work group, members take interest in one another." Respondents indicated their agreement with these items using a 5-point response format (1 = *strongly disagree*, 5 = *strongly agree*). Alphas for the paper-and-pencil and the Web-based versions, respectively, were .92 and .93.

Collective goal commitment. Six items from Hollenbeck, Klein, O'Leary, and Wright (1989) were included to gauge respondents' goal commitment. The items were slightly modified from their original form. Whereas Hollenbeck et al.'s original items referred to the individual (e.g., "I am strongly committed to pursuing this goal"), we used a referent-shift (Chan, 1998) so that items referred to the collective ("In my work group, people are strongly committed to pursuing our goals"). Respondents indicated their agreement with these items using a 5-point response format (1 = *strongly disagree*, 5 = *strongly agree*). Alpha for the paper-and-pencil version was .85; alpha for the Web-based version was .86.

Procedure

Respondents completed all four study measures in the same sequence across both administration modes. The Web-based survey was pilot tested to identify potential incompatibilities

among operating systems, Web browsers, and so on. After the pilot test was completed, a date was set for survey administration. An electronic message was sent from the CEO to all employees who had a company e-mail account. The message contained a brief salutation, a short description of the company's interest in the research project, the deadline for completing and returning the survey (11 days later), and a link to the Web portal where the survey was posted. The message also explained that participation was voluntary, and anonymity was guaranteed. Because this was the first time employees were invited by the company to complete a survey online, a description of how data were transmitted and stored on a third-party's computer server (affiliated with the university conducting the study) was also provided. One week later, a reminder was sent from the CEO's e-mail account. This reminder thanked everyone who had already participated, provided the Web location where the survey was posted, and noted that the survey deadline was approaching.

The company was accustomed to communicating in various languages and had systems in place for doing so (e.g., employees' e-mail addresses were grouped by country of residence). This capability allowed for the CEO's invitation and reminder messages to be specifically tailored to the predominant language(s) spoken in each home country. Whereas employees located in the United States received e-mail in English, employees in Switzerland, for example, received e-mail in German, French, Italian, and English.

The paper-and-pencil surveys were delivered via internal company mail. In most locations, several versions of the survey (e.g., the English version and one to three other languages) were made available so that employees could choose the language in which they were most comfortable conversing. The company announced the survey through internal memoranda, and local managers were encouraged to ensure that all employees were aware of the survey's purpose and availability. With regard to the format of the paper-and-pencil survey, the software used to develop the Web-based survey provided us with the capacity to also develop a portable document formatted (pdf) file that printed a look-alike print copy. When completing a traditional paper-and-pencil survey, respondents can preview, review, skip, and even change answers if desired. Likewise, in the present study, respondents who completed the Web-based survey were also allowed to preview, review, skip, and change their answers. The look-alike appearance and identical operational features were specifically intended to control for confounding effects due to differences in administrative procedures.

To complete the study, respondents to the Web survey were instructed to click a button labeled *Submit*. Respondents completing paper-and-pencil surveys were instructed to mail them to the university address of the primary author.

Analyses

For each of the aforementioned measures, tests of measurement equivalence were conducted in seven sequential phases. In each phase, we employed MGCFA to compare the covariance matrices of individual measure items. Missing data accounted for roughly 3% to 4% of the total number of item ($n = 38$) responses. Congruent with Stanton (1998), the average number of missing items per paper-and-pencil respondent ($M = 1.83$) was slightly higher than that of Web respondents ($M = 1.13$). To impute missing values, we used

the full-information maximum likelihood (FIML) technique. FIML is superior to other imputation techniques as it gives unbiased estimates of means, variances, and other parameters (Arbuckle, 1996; Byrne, 2001; Wothke, 2000).

Like standard confirmatory factor analysis, overall model fit in MGCFA is commonly evaluated for statistical significance in which a nonsignificant chi-square indicates a good fit. Because of a functional dependence on sample size, for large samples, the chi-square statistic provides a sensitive statistical test but not a practical test of model fit. Consequently, a second approach is to use practical fit indices as an alternative to chi-square (Hu & Bentler, 1998). Accordingly, we employed three additional indices for assessing overall model fit: (a) the comparative fit index (CFI), which compares how much better an implied model fits as compared to a null model; (b) the Tucker-Lewis index (TLI), which contains a penalty for lack of parsimony; and (c) the root mean square error of approximation (RMSEA), which adjusts for both sample size and degrees of freedom. With regard to RMSEA, we also created confidence intervals (LO90 and HI90) as recommended by Byrne (2001).

In testing measurement equivalence, we were interested in making cross-mode comparisons (i.e., paper and pencil vs. Web based) with regard to the equality of estimated parameters of two nested models. This procedure involves testing the fit of a series of increasingly restrictive models against a baseline model. To determine the degree of equivalence, previous research has employed the likelihood ratio test (also known as the chi-square difference test). Chi-square differences between nested models are, however, sample size dependent (Brannick, 1995). Similarly, the chi-square statistic is known to be an overly sensitive index of model fit for models with numerous constrained parameters fitted on large samples (Marsh, Balla, & McDonald, 1988). Nonetheless, because no standard exists for comparing changes in practical fit indices when constraints are added (Vandenberg & Lance, 2000), researchers using MGCFA have had to rely on the likelihood ratio test to assess whether a constrained model fits obtained data less well than a less-constrained model does. To address this limitation, Cheung and Rensvold (2002) conducted a Monte Carlo simulation to assess differences in practical goodness-of-fit indices under the null hypothesis of measurement equivalence. Their findings suggested most changes in practical goodness-of-fit indices (ΔGFIs) are dependent on both sample size and model complexity and are also correlated with overall fit measures. The ΔCFI metric, however, was found to be a robust fit statistic when testing MGCFA models. Furthermore, they found that a critical ΔCFI value ≤ -0.01 indicates that a null hypothesis of equivalence should not be rejected. Thus, we tested our equivalence hypotheses using this critical value.

Phase 1 tested for configural equivalence to determine if the patterns of fixed and free factor loadings were comparable across the paper-and-pencil and Web-based surveys (Vandenberg, 2002). Given that configural equivalence exists, the same number of underlying factors would emerge. Differing factor structures across administration modes would be evidence that respondents to the two survey forms used different conceptual frames of reference. As previously noted, a failure to support configural equivalence argues against further tests of measurement invariance (Vandenberg & Lance, 2000). Phase 2 tested for metric invariance to determine if factor loadings for items were equal across administration modes (Vandenberg, 2002). In Phase 2, the model tested as a part of Phase 1 was constrained by setting the factor loadings to be equal for both administration modes. A loss in fit, as indicated

by the ΔCFI between the restricted model and the model tested in Phase 1, would support rejecting a null hypothesis of invariant factor loadings (Cheung & Rensvold, 2002).

Phase 3 examined scalar invariance to determine whether, in addition to invariant factor loadings, the vectors of item intercepts were also equal across administration modes (Vandenberg, 2002). Cheung and Rensvold (2002) have indicated that scalar equivalence is a prerequisite for the comparison of latent means because it suggests that underlying item-response scales have the same intervals and zero points. Accordingly, they have contended that if scalar equivalence is not supported, any subsequent mean comparisons remain ambiguous. As noted, however, others have argued that latent-mean comparisons can still be made under partial scalar equivalence (e.g., Byrne et al., 1989; Millsap & Kwok, 2004; Steenkamp & Baumgartner, 1998). Following this latter rationale, if at least partial intercept equivalence were present, we compared latent means across administration modes.

If the increasingly restrictive models in Phases 2 and 3 showed no deterioration in fit (i.e., ΔCFI) when compared to the baseline model tested in Phase 1, we moved to Phase 4. In this phase, we tested the null hypothesis that like items' residual error variances would be equivalent across administration modes. Therefore, in Phase 4, the factor structures, factor loading parameters, intercepts, and error variances were constrained to be equal for both modes. Because residual variance cannot be attributed to the variance associated with an underlying latent variable, equality of error variances across administration modes determines if survey items tap their associated latent variables with the same degree of measurement error (Cheung & Rensvold, 2002). Unfamiliarity with scoring formats, differences in vocabulary, idioms, grammar, and reactivity resulting from different administration modes have all been suggested as sources of residual nonequivalence (Cheung & Rensvold, 2002; Potosky & Bobko, 1997; Stanton, 1998). Consistent with the earlier phases, a change in the CFI of ≥ -0.01 was taken as evidence that a null hypothesis of invariant residual variances should be rejected.

The three remaining data analysis phases involved testing for construct-level equivalence. Phases 5 and 6 addressed relational equivalence for all four study measures. Cheung and Rensvold (2002) noted that when one's purpose is to relate a theoretical construct (in our case, transformational leadership) to other constructs (viz., collective efficacy, work-group cohesiveness, and collective goal commitment) in a nomological network, latent-construct variance equivalence must be present. Given support for construct-variance equivalence (i.e., Phase 5), we tested for equality between network correlations in Phase 6 by adding an additional constraint of equal covariances. The finding that both construct variances and covariances are invariant across administration modes indicates that the correlations between latent constructs are equal for both modes (e.g., Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000). Steenkamp and Baumgartner (1998) have explained, however, that when measures of association between variables are compared across groups, measurement reliabilities should "be about the same so that measurement artifacts do not bias the substantive conclusions" (p. 82). Measurement error equivalence is an indicator of a measure's reliability (Vandenberg & Lance, 2000), and therefore, the equivalence procedure tested in phase 4 was a prerequisite for testing the relational equivalence of all four study measures.

Finally, in Phase 7, we tested the null hypothesis of equivalent latent means across administration modes. It has been suggested that latent-mean differences are more valid than simple analysis of variance (ANOVA) or *t* tests because observed mean differences

Table 2
Skewness, Kurtosis, Reliabilities, and Intercorrelations for All Study Variables

	Web-Based Mode			Paper-and-Pencil Mode			*r*			
Measure	Skewness	Kurtosis	α	Skewness	Kurtosis	α	1	2	3	4
Transformational leadership	–.45	–.30	.96	–.39	–.53	.96	—	.37**	.39**	.45**
Collective efficacy	–.30	.08	.77	–.38	.02	.74	.30**	—	.68**	.67**
Work-group cohesiveness	–.71	.30	.93	–.70	.12	.92	.42**	.70**	—	.64**
Collective goal commitment	–.69	.67	.86	–.52	.02	.85	.41**	.69**	.71**	—

Note: Correlations above the diagonal are for the Web-based mode (*n*s = 4,117 to 4,164), whereas those below the diagonal are for the paper-and-pencil mode (*n*s = 633 to 648).
**$p < .01$.

cannot be attributable to a lack of equivalence (e.g., Ployhart & Oswald, 2004; Vandenberg & Lance, 2000). In the present study, the procedure outlined by Byrne (2001) for latent-mean comparisons was followed. As part of this procedure, it was necessary to allow latent-variable means to be freely estimated for one administration mode but to be constrained equal to zero for the other, which thus served as a "reference group." Determination of which mode would serve as the reference group is purely arbitrary (Byrne, 2001). We chose the paper-and-pencil mode to serve this purpose. Thus, the latent means for the Web-based mode were estimated. Significantly higher mean values would indicate that the Web-based mode reported higher mean ratings on a given measure. In determining significant differences, the critical ratio follows an approximately normal distribution, with values greater than ±1.96 indicating statistical significance ($p \leq .05$; Arbuckle & Wothke, 1999).

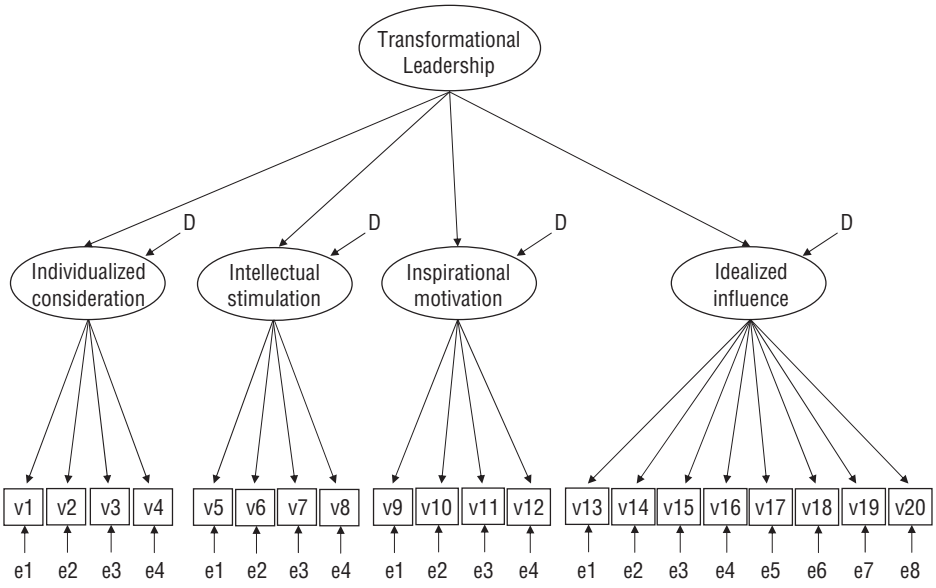
Results

In addition to skewness, kurtosis, and coefficient alphas for the observed variables, the intercorrelations among all study variables for the Web-based and paper-and-pencil samples are reported in Table 2. Of note are the variables' internal consistency estimates and the association between transformational leadership and the three outcomes. The coefficient alphas and intercorrelations among the focal variables appear similar across administration modes.

The Factor Structure of Bass and Avolio's Transformational Leadership Measure

Within each subsample (i.e., Web-based and paper-and-pencil), we first investigated the factor structure of transformational leadership by comparing two measurement models. In previous studies, intercorrelations among transformational dimensions were generally high and positive (Lowe, Kroeck, & Sivasubramaniam, 1996). In such situations, whether these dimensions are empirically distinguishable has been questioned (e.g., Bycio, Hackett, &

Figure 1
Second-Order Measurement Model for Transformational Leadership



Note: D = disturbance; v = variable; e = residual error term.

Allen, 1995; Carless, 1998). Accordingly, the first two models fitted the Web-based data (Model 1a) and paper-and-pencil data (Model 1b) to load all 20 transformational items on a global transformational leadership construct. Results indicated that the global leadership factor was an acceptable fit to both the Web-based ($\chi^2 = 6126.11$, $df = 170$, $p < .001$, CFI = .975, TLI = .969, RMSEA = .091, LO90 = .089 and HI90 = .093) and the paper-and-pencil data ($\chi^2 = 904.04$, $df = 170$, $p < .001$, CFI = .978, TLI = .973, RMSEA = .081, LO90 = .076 and HI90 = .086).

In Models 2a and 2b, we fitted a single higher order model to the data (see Figure 1). In his original validation work, Bass (1985) discussed the importance of considering higher order models. Building from this base, Avolio, Bass, and Jung (1999) reported that a second-order model successfully explained the covariation among first-order factors representing the transformational dimensions. In support, others have reported results revealing a single, higher order transformational leadership factor (e.g., Carless, 1998; Den Hartog, Van Muijen, & Koopman, 1997). As shown in Figure 1, items in Models 2a and 2b were fixed to load on their respective factors (individualized consideration, intellectual stimulation, inspirational motivation, and idealized influence), and a higher order transformational leadership factor was also modeled to explain the covariation among latent first-order factors. Results indicated that the higher order model fit the paper-and-pencil data (Model 2b: $\chi^2 = 792.82$, $df = 166$, $p < .001$, CFI = .981, TLI = .976, RMSEA = .075, LO90 = .070 and HI90 = .081); however, in the Web-based sample, an improper solution, resulting from a negative residual estimate, was obtained. Such improper solutions are frequently encountered in practice, and

numerous approaches have been suggested for handling such offending estimates (see Chen, Bollen, Paxton, Curran, & Kirby, 2001; Lawley & Maxwell, 1971; Dillon, Kumar, & Mulani, 1987). When fitting hierarchical models, Byrne (2001) has noted the importance of inspecting the higher order portion of a model using the critical ratio difference (CRDIFF) method. Indeed, the critical ratios of differences for the residual variances were found to be, at times, statistically nonsignificant in the samples (one of six comparisons in the Web-based sample; three of six comparisons in the paper-and-pencil sample). Based on these findings, we followed Byrne's recommendation and constrained the residual variances to be equal. We then fit the revised model (Model 2c) to our data. There was no indication of an improper solution, and the practical fit indices indicated a good fit ($\chi^2 = 4834.45$, $df = 169$, $p < .001$, CFI = .981, TLI = .976, RMSEA = .081, LO90 = .079 and HI90 = .083).

With the models independently fitted to the data, we then used the chi-square difference test to determine the best fitting measurement model for the transformational leadership measure. Compared to Model 1a, the revised Model 2c exhibited a better fit, $\Delta\chi^2 = 1291.66$, $\Delta df = 1$, $p < .001$, and all practical fit indices improved.[1] Likewise, results for the paper-and-pencil data (Model 1b – Model 2b) indicated that the single higher order model was a better fit, $\Delta\chi^2 = 111.22$, $\Delta df = 4$, $p < .001$, and the practical fit indices remained virtually unchanged. Supporting earlier studies, a second-order hierarchical factor structure provided the best fit to the data.

Hypothesis Testing for Web-Based Versus Paper-and-Pencil Administration Modes

It is generally recommended that researchers independently inspect each measure's factor structure (i.e., configural equivalence) to determine the most appropriate baseline measurement model (Byrne, 2001). We first examined the configural equivalence for the higher order transformational leadership model across the administration modes. Each practical fit index was within acceptable guidelines (CFI = .980, TLI = .976, and RMSEA = .057, LO90 = .055 and HI90 = .058), supporting configural equivalence with respect to transformational leadership across administration modes (see Table 3). Independent baseline analyses were similarly conducted for collective efficacy, cohesion, and collective goal commitment. Results shown in Table 3 provide support for each measure's configural invariance. All practical fit indices were within acceptable ranges. For collective efficacy, the CFI = .990 and TLI = .978 were high, and the RMSEA = .093 (LO90 = .087 and HI90 = .099) was slightly below the .10 guideline. Fit indices for cohesion (CFI = .995, TLI = .989, and RMSEA = .072, LO90 = .067 and HI90 = .078) and collective goal commitment (CFI = .995, TLI = .989, and RMSEA = .070, LO90 = .065 and HI90 = .076) also suggested a good fit.

Whereas the prior analyses provided initial support for each measure's configural equivalence when examined independently, a more stringent configural equivalence test would incorporate all four measures into one measurement model and simultaneously examine the model's fit to the Web-based and paper-and-pencil data. We therefore developed a measurement model whereby each measure's items were fixed to load on its respective latent factor (no item cross-loadings were permitted), and the four latent factors (higher order transformational leadership model, collective efficacy, work-group cohesiveness, and

Table 3
Goodness-of-Fit Tests for the Multigroup Confirmatory Factor Analyses

Instrument	χ^2	*df*	CFI	TLI	RMSEA	RMSEA 90% Confidence Intervals
Separate measurement models						
Transformational leadership	5,691.97***	339	.980	.976	.057	.055-.058
Collective efficacy	781.88***	18	.990	.978	.093	.087-.099
Work-group cohesiveness	481.88***	18	.995	.989	.072	.067-.078
Collective goal commitment	454.19***	18	.995	.989	.070	.065-.076
Phase 1: Configural equivalence						
Equal factor structures	10,736.54***	1317	.983	.981	.038	.038-.039
Phase 2: Metric equivalence						
Equal factor loadings	10,823.21***	1352	.983	.981	.038	.037-.038
Phase 3: Scalar equivalence						
Equal intercepts	11,454.71***	1390	.982	.981	.038	.038-.039
Phase 4: Residual equivalence						
Equal uniquenesses	12,092.53***	1428	.981	.980	.039	.038-.040
Phase 5: Latent-construct equivalence						
Equal latent variances	12,101.55***	1431	.981	.980	.039	.038-.040
Phase 6: Covariance equivalence						
Equal covariances	12,136.45***	1437	.981	.980	.039	.038-.040

Note: CFI = comparative fit index; TLI = Tucker-Lewis index; RMSEA = root mean error of approximation.
***$p < .001$.

collective goal commitment) were allowed to covary. Consequently, this four-factor measurement model was used as the baseline model in the subsequent tests of measurement equivalence.

Phase 1 tested the null hypothesis of an equal number of factors across administration modes for the study's measures. As shown in Table 3, the four-factor model was an excellent fit to the data, $\chi^2 = 10{,}736.5$, $df = 1317$, $p < .001$, CFI = .983, TLI = .981, RMSEA = .038, LO90 = .038 and HI90 = .039. Given that the fit indices indicated acceptable fit, we failed to reject the null hypothesis of equal factor structures and therefore proceeded to Phase 2.

Phase 2 tested all four measures across administration modes for equal factor loadings. Because of the single higher order transformational leadership factor, the equality constraint was imposed on the second-order path coefficients as well. As described earlier, we used Cheung and Rensvold's (2002) suggested ΔCFI of ≤ -0.01 as an indicator that the constrained model tested in Phase 2 adequately fit the data. Fit indices are reported in Table 3. Inspecting the ΔCFI compared to the baseline model in Phase 1 indicated the CFI value did not change, despite the added constraints. Similarly, the TLI and RMSEA values were also similar to those reported in Phase 1. Taken together, these results support the conclusion that the factor loadings associated with their respective measures were invariant across the Web-based and paper-and-pencil modes. On this basis, we failed to reject the null hypothesis of equal factor loadings and proceeded to Phase 3 of our analyses.

In Phase 3, we tested scalar invariance across administration modes by adding the additional constraint of equal item intercepts to the already restricted model. As reported in Table 3, the CFI and TLI were .982 and .981, respectively, and the RMSEA value continued to suggest a good fit (.038). Inspecting the ΔCFI compared to the baseline model in Phase 1, the CFI change of −0.001 was less than the critical threshold as recommended by Cheung and Rensvold (2002). Consequently, the intercepts of like items were equal across administration modes when regressed on their associated latent factors. These findings suggest that all four measures exhibited scalar invariance, a necessary precondition for the unambiguous comparisons of latent means.

Although testing for invariance in measurement error is considered by some to be an overly restrictive test (Cheung & Rensvold, 1999; Little, 1997), we also applied this additional constraint to our measurement model. Thus, in Phase 4, we tested for invariance in measurement error across the items comprising each study measure. In addition to equal factor loadings and intercepts, the error variance associated with each item in a given measure was constrained to be equal across administration modes. These results are reported in Table 3. All practical fit indices indicated the Phase 4 model exhibited a good fit. Most important, the change in the CFI (ΔCFI = −0.002) indicated there was no deterioration in model fit despite increasingly restrictive equality constraints. Consequently, we failed to reject the null hypothesis of equal item residual variance. Having found that the error variances associated with individual items were equivalent across modes (in addition to scalar equivalence in Phase 3), we concluded that all four measures were reliable across both administration modes.

In Phase 5, we tested the latent construct invariance of each measure across administration modes by adding the constraint of equal construct variances to the already restricted model of invariant residual variances. As reported in Table 3, the practical fit indices suggested the model continued to provide a good fit. An inspection of the ΔCFI compared to the baseline model in Phase 1 (−0.002) demonstrated that there was no model degradation despite the additional constraint of equal construct variances. These findings suggest that the range of responses to each measure's items was comparable across Web-based and paper-and-pencil modes. Thus, construct variances were invariant across administration modes.

With equivalent construct variances established, in Phase 6, we applied the additional constraint of invariant covariances to the already highly restricted measurement model. Results are reported in Table 3. Practical fit indices indicated that there was no serious model deterioration despite equality constraints placed on the construct covariances. Also of importance, the ΔCFI (−0.002) compared to the configural model estimated in Phase 1 was less than the critical value recommended by Cheung and Rensvold (2002). Considered collectively, findings observed in Phases 5 to 6 demonstrated that both construct variances and covariances were invariant across administration modes. Hence, the latent correlations among the four focal constructs were also equivalent across modes. Based on this result, we concluded that relational equivalence existed across the Web-based and paper-and-pencil surveys.

In sum, the present findings provided evidence for the equivalence of the overall measurement structure (Phase 1), factor loadings (Phase 2), item intercepts (Phase 3), item

uniquenesses or residuals (Phase 4), factor variances (Phase 5), and factor covariances (Phase 6) across administration modes. Nevertheless, we must again note that our conclusions were based on the ΔCFI statistic as recommended by Cheung and Rensvold (2002). Although reported to be a robust index for testing the multigroup invariance of CFA models (Cheung & Rensvold, 2002), an alternative explanation for our finding full equivalence is that the ΔCFI possesses low power as a model-comparison procedure. In the present application, however, low power did not seem to be an applicable concern. As shown in Table 3, the CFIs and TLIs remained virtually unchanged, and the RMSEA 90% confidence intervals were overlapping across models. Moreover, the standardized regression weights (factor loadings) obtained for the full equivalence model (Phase 6) were similar to the observed factor loadings for the freely estimated model (Phase 1). Furthermore, as illustrated in Table 4, the differences (Phase $6_{\text{fully constrained model}}$ – Phase $1_{\text{freely estimated model}}$) between the standardized parameter estimates for the Web-based measures (|0.006|) and the paper-and-pencil measures (|0.026|) were small.

Because full equivalence strictly held for all four measures, valid comparisons of latent-factor means were conducted in Phase 7. To determine if mean differences existed, latent-mean comparisons were estimated using the Phase 5 model, but the latent means for the paper-and-pencil sample were constrained to zero. As shown in Table 5, there were differences between the paper-and-pencil and Web-based latent means. For three of four cases, the Web-based sample reported significantly higher measure scores. Specifically, higher measure scores were obtained for transformational leadership ($\text{mean}_{\text{diff}} = 0.162$, $p < .01$), work-group cohesiveness ($\text{mean}_{\text{diff}} = 0.093$), $p < .01$), and collective goal commitment ($\text{mean}_{\text{diff}} = 0.083$, $p < .05$). In contrast, the paper-and-pencil sample expressed a higher level of collective efficacy ($\text{mean}_{\text{diff}} = |0.065|$, $p < .01$) than did the Web-based sample. On the whole, the mean differences were not large; the largest mean difference was found for transformational leadership (|0.162|, $p < .01$). Finally, a review of the practical fit indices indicated the model fit the data ($\chi^2 = 12{,}023.2$, $df = 1427$, $p < .001$, CFI = .981, TLI = .980, RMSEA = .039, LO90 = .038 and HI90 = .040). Given that the practical fit indices were within acceptable guidelines, we are confident in our interpretation of the latent-mean estimates (see Byrne, 2001, p. 241).

For comparison purposes, we also estimated latent means within the context of a model that does not assume full equivalence (as incorrectly assuming equivalence may bias the test of latent means). When estimating latent means, the attainment of an overidentified model is only possible with the imposition of additional model constraints (see Byrne, 2001, p. 229). As a consequence, we conducted a second latent-means analysis using the scalar equivalence model. As shown at the bottom of Table 5, the latent-mean differences were nearly identical to those reported for the full equivalence model.

Confound Analyses

Because the Web-based and paper-and-pencil samples were drawn from multiple countries and surveys were presented in different languages, we assessed each measure's equivalence across countries and languages to rule out the possibility of cross-national confounds. The two samples also appeared to systematically differ with respect to job level.

Table 4
Standardized Factor Loadings Contrasting the Phase 1 Versus Phase 6 Measurement Equivalence Tests

	Phase 1: Configural Equivalence[a]		Phase 6: Covariance Equivalence[b]	
Instrument	Web-Based Sample	Paper-and-Pencil Sample	Web-Based Sample	Paper-and-Pencil Sample
TL-MLQ				
Second-order factor loadings				
Individualized consideration	.938	.964	.960 (–.022)	.960 (.004)
Intellectual stimulation	.936	.953	.941 (–.005)	.941 (.012)
Inspirational motivation	.960	.951	.939 (.021)	.939 (.012)
Idealized influence	.971	.970	.971 (.000)	.971 (–.001)
First-order factor loadings				
Individualized consideration				
Item 1	.736	.760	.741 (–.005)	.741 (.019)
Item 2	.652	.648	.654 (–.002)	.654 (–.006)
Item 3	.703	.652	.698 (.005)	.698 (–.046)
Item 4	.844	.840	.844 (.000)	.844 (–.004)
Intellectual stimulation				
Item 1	.697	.718	.702 (–.005)	.702 (.016)
Item 2	.754	.754	.753 (.001)	.753 (.001)
Item 3	.798	.802	.799 (–.001)	.799 (.003)
Item 4	.818	.798	.814 (.004)	.814 (–.016)
Inspirational motivation				
Item 1	.674	.677	.673 (.001)	.673 (.004)
Item 2	.809	.723	.796 (.013)	.796 (–.073)
Item 3	.776	.787	.778 (–.002)	.778 (.009)
Item 4	.813	.771	.804 (.009)	.804 (–.003)
Idealized influence (attributed and behavior)				
Item 1	.826	.786	.821 (.005)	.821 (–.035)
Item 2	.761	.737	.759 (.002)	.759 (–.022)
Item 3	.835	.827	.835 (.000)	.835 (–.008)
Item 4	.775	.771	.773 (.002)	.773 (–.002)
Item 5	.758	.841	.772 (–.014)	.772 (.069)
Item 6	.742	.669	.731 (.011)	.731 (–.062)
Item 7	.695	.629	.687 (.008)	.687 (–.058)
Item 8	.648	.656	.649 (–.001)	.649 (.007)
Collective efficacy				
Item 1	.534	.524	.529 (.005)	.529 (–.005)
Item 2	.451	.415	.444 (.007)	.444 (–.029)
Item 3	.730	.742	.732 (–.002)	.732 (.010)
Item 4	.776	.787	.771 (.005)	.771 (.016)
Item 5	.542	.401	.519 (.023)	.519 (–.118)
Item 6	.700	.637	.690 (.010)	.690 (–.053)

(continued)

Table 4 (continued)

Instrument	Phase 1: Configural Equivalence[a]		Phase 6: Covariance Equivalence[b]	
	Web-Based Sample	Paper-and-Pencil Sample	Web-Based Sample	Paper-and-Pencil Sample
Work-group cohesiveness				
Item 1	.749	.662	.736 (.013)	.736 (–.074)
Item 2	.821	.799	.818 (.003)	.818 (–.019)
Item 3	.862	.842	.855 (.007)	.855 (–.013)
Item 4	.846	.851	.847 (–.001)	.847 (.004)
Item 5	.865	.870	.866 (–.001)	.866 (.004)
Item 6	.808	.819	.810 (–.002)	.810 (.009)
Collective goal commitment				
Item 1	.785	.816	.789 (–.004)	.789 (.027)
Item 2	.646	.700	.652 (–.006)	.652 (.048)
Item 3	.639	.538	.624 (.015)	.624 (–.086)
Item 4	.752	.694	.744 (.008)	.744 (–.050)
Item 5	.695	.691	.694 (.001)	.694 (–.003)
Item 6	.735	.732	.736 (–.001)	.736 (–.004)

Note: TL-MLQ = transformational leadership measure. Numbers within parentheses reflect the difference between the freely estimated factor loadings and their respective loadings for the completely constrained model. All factor loadings are significant, $p < .001$.

a. All parameters were freely estimated.

b. Complete equivalence model. All parameters were constrained to equality.

Table 5
Results of Latent Mean Difference Tests

Measure	Mean difference[b]	Standard Error	Critical Ratio
Critical values[a]			
Transformational leadership	0.162	0.031	5.29**
Collective efficacy	–0.065	0.023	–2.80**
Work-group cohesiveness	0.093	0.033	2.84**
Collective goal commitment	0.083	0.033	2.52*
Critical values[c]			
Transformational leadership	0.162	0.032	5.11**
Collective efficacy	–0.073	0.024	–3.00**
Work-group cohesiveness	0.094	0.035	2.68**
Collective goal commitment	0.079	0.034	2.30*

a. Mean differences for complete equivalence model (Phase 6).

b. Paper-and-pencil mode was used as the referent; values reflect mean differences for the Web-based mode.

c. Mean differences for scalar equivalence model (Phase 3).

*$p < .05$. **$p < .01$.

The paper-and-pencil sample included employees who did not have a company e-mail address, and therefore the sample was primarily composed of lower level, nonmanagerial employees. As a result, we conducted a series of MGCFAs to evaluate the equivalence of responses across these various groups. To the extent that measurement equivalence holds for the measures across country, language, and job-level groups, any concerns that involve sample differences confounding the administration-mode results should be lessened.

The organizational samples varied in size, particularly when considering the differing sample sizes across countries. Because MGCFA generally requires moderate to large sample sizes, we limited the analyses to only those samples that exceeded 100 participants. When evaluating cross-national or cross-language measurement equivalence, previous researchers have suggested that equivalence is supported when factor loadings (metric equivalence) are found equivalent across groups (Ryan et al., 1999). Therefore, based on the ΔCFI critical value (Cheung & Rensvold, 2002), equal factor loadings across groups were used as the determinant in the present study to support cross-national and cross-language equivalence.

We first tested cross-national equivalence using data from 13 countries: Australia ($n = 106$), Canada ($n = 128$), Croatia ($n = 414$), Denmark ($n = 148$), France ($n = 508$), Italy ($n = 186$), Poland ($n = 210$), Romania ($n = 117$), Spain ($n = 101$), Sweden ($n = 253$), Switzerland ($n = 753$), United Kingdom ($n = 328$), and the United States ($n = 682$). Results indicated there was no model degradation (ΔCFI value was −0.001) despite the added equality constraint on the factor loadings when the constrained model was compared to the baseline model of configural equivalence. Next, equivalence comparisons across 12 languages were conducted. Included in this analysis were 12 language versions of the four measures: Croatian ($n = 417$), Czech ($n = 104$), Danish ($n = 146$), English ($n = 1{,}727$), French ($n = 602$), German ($n = 684$), Italian ($n = 196$), Polish ($n = 214$), Portuguese ($n = 102$), Romanian ($n = 117$), Spanish ($n = 174$), and Swedish ($n = 258$). In comparing the configural (i.e., baseline) and constrained model (i.e., metric equivalence), there was no indication of model degradation (ΔCFI value was −0.002), thus fully supporting metric equivalence across languages for each of the four measures.

Following the recommendation of Ryan et al. (1999), we then conducted a cross-national, cross-language comparison using the five largest samples, which included the United States (English language only, $n = 672$), the United Kingdom (English language only, $n = 326$), France (French language only, $n = 490$), Croatia (Croatian language only, $n = 409$), and Sweden (Swedish language only, $n = 250$) samples. We chose not to include Switzerland in this analysis because the country does not have one distinct national language; depending on the geographic location, German, Italian, or French is the language most used. Consistent with the previous findings, metric equivalence was fully supported for the measures (ΔCFI value was −0.002) across the five cross-national, cross-language groups. Finally, in that the paper-and-pencil sample was composed primarily of nonmanagerial employees, only four groups were included in this analysis: upper-level managers (Web-based version, $n = 500$), line managers and team leaders (Web-based version, $n = 1{,}246$), nonmanagerial employees (Web-based version, $n = 2{,}498$), and nonmanagerial employees (paper-and-pencil version, $n = 589$). Once again, the measures exhibited metric equivalence

across the job-level groups (ΔCFI = 0.000). In sum, these findings support measurement equivalence across country, language, and job level.

Discussion

The purpose of the present investigation was to examine, within an international context, various measurement equivalence issues related to paper-and-pencil and Web-based versions of Bass and Avolio's (2000) TL-MLQ, the most frequently used measure of transformational leadership. Moreover, as an extension of previous measurement invariance research, we sought to examine whether administration mode affects whether TL-MLQ responses have equivalent relations with three theoretically related workplace constructs (viz., collective efficacy, work-group cohesiveness, and collective goal commitment).

In line with this purpose, our findings support the need to focus on whether the Web-based administration of well-developed paper-and-pencil measures can result in measurement invariance. Because we employed a series of nested statistical tests using MGCFA, we may more confidently conclude that the observed results were valid and not due to measurement artifacts. Consequently, several conclusions are worth noting.

First, the underlying factor structures in our study measures were the same across administration modes. Thus, there was no evidence that paper-and-pencil and Web-based versions of the same measures evoked different conceptual frames of reference. Combined with the finding that factor loadings across administration modes were equivalent, we conclude that the measures tested displayed "strong factorial invariance" (Vandenberg, 2002). Furthermore, respondents completed the measures in one of 17 languages. When working with so many different languages, there is always a concern for translation error (see, e.g., Mullen, 1995, pp. 575-576). Had translation error been influential in the present study, however, configural and metric invariance would have been absent (Cheung & Rensvold, 2002; Little, 1997).

Second, no evidence of unequal item intercepts was found in any of our study measures. As a result, the observed scalar invariance suggests that these measures possessed equal intervals and zero points. Therefore, each measure was isometrically calibrated across administration modes. Third, it has been suggested that the use of different scoring formats and, more recently, varying administration modes, might be a source of measurement error (Mullen, 1995; Stanton, 1998). Despite varying dimensionality, length, and use of differing scale anchors (i.e., frequency and agreement), there was no indication that the unique error variances associated with the items comprising any of the four study variables differed across mode. Consequently, the present findings add to a growing body of research that suggests online instruments do not unduly influence the psychometric properties of well-developed paper-and-pencil measures (Gosling et al., 2004; Herrero & Meneses, in press; Potosky & Bobko, 1997).

One important methodological issue was our decision to follow Cheung and Rensvold's (2002) recommendation to use the ΔCFI index to detect a lack of invariance.[2] Despite various factors that could be expected to contribute to a lack of invariance, based on the ΔCFI guideline, we concluded that the measures were fully equivalent across administration modes. Stated differently, some might question ΔCFI's ability to detect a lack of invariance

when in fact it exists, or what Vandenberg (2002, pp. 142-147) has labeled the "sensitivity issue." The issue of sensitivity concerning the ΔCFI is notable given that Cheung and Rensvold's (2002) study examined the performance of goodness-of-fit indices with respect to Type I error but not Type II error. This is an understandable concern, but there were two main reasons for not investigating Type II errors (G. Cheung, personal communication, August 23, 2005). First, there are no commonly accepted measures of effect sizes for differences in estimated parameters when using covariance structure analysis. Therefore, a meaningful simulation study of Type II errors is impossible unless it is based on arbitrary differences in one or more estimated parameters, resulting in study findings that are not generalizable to other settings. Second, a major issue associated with Type II errors is the overpower of the likelihood ratio test used to determine the degree of equivalence. By contrast, there is little concern about inadequate power in applying the likelihood ratio test.

Concerning the ΔCFI's sensitivity with regard to the current research, two additional points merit mention. Consider, for example, the reported RMSEA values for each of the measurement equivalence hypotheses. As shown in Table 3 (Phases 1-6), the 90% confidence intervals consistently overlapped, indicating that there was no significant model deterioration despite the increasingly restrictive equality constraints that were imposed on our measurement model. These results provide additional evidence that the four measures tested were equivalent across administration modes, as likewise suggested by the ΔCFI values. Second, using the Social Sciences Citation Index (http://www.isinet.com/products/citation/wos/) database, we located more than 30 articles that referenced Cheung and Rensvold's (2002) study, with a handful of these employing the ΔCFI guideline to detect a lack of measurement equivalence. In several of these articles, researchers reported failing to find support for null hypotheses of measurement equivalence because the ΔCFI exceeded the −0.01 threshold (e.g., Asci, Eklund, Whitehead, Kirazci, & Koca, 2005; French & Tait, 2004; Lievens & Anseel, 2004; Mantzicopoulos, French, & Maller, 2004; Ployhart, Wiechmann, Schmitt, Sacco, & Rogg, 2003). Thus, whereas we acknowledge that it may be premature to definitively conclude that the ΔCFI is the index for establishing measurement equivalence, we have no reason to doubt its sensitivity in detecting a lack of invariance.

A final point worth noting is that latent-mean scores for three of our study measures differed significantly across administration modes. With the exception of collective efficacy, computer-based scores were higher than their paper-and-pencil counterparts. When past studies have found such differences, paper-and-pencil measures have usually yielded relatively more positive attitudes (e.g., Eaton & Struthers, 2002; Stanton, 1998). We believe there are three possible explanations for our contradictory finding.

One possible explanation is that the Web-based application prompted respondents to present themselves in a socially desirable manner, thus creating a general response set. Richman, Kiesler, Weisband, and Drasgow (1999) referred to social-desirability distortion as "the tendency by respondents, under some conditions and modes of data-collection, to answer questions in a more socially desirable direction than they would under other conditions or modes of administration" (p. 755). For years, it was believed that the anonymous, impersonal, and nonjudgmental nature of Web-based surveys would reduce desirable responding; past findings, however, are mixed (e.g., Lautenschlager & Flaherty, 1990; Rosenfeld, Booth-Kewley, Edwards, & Thomas, 1996; Whitener & Klein, 1995). In their

meta-analysis of the literature, Richman et al. (1999) reported that in certain situations (e.g., expectations on how data were to be used or being told that responses would be kept in a data file), social-desirability distortion was greater in Web-based as compared to paper-and-pencil surveys. Others have similarly noted that respondents' beliefs concerning the trustworthiness and reliability of Web-based data are increasingly becoming adversely affected by the publicity given to Internet privacy issues (Simsek & Veiga, 2001; Stanton & Rogelberg, 2001). Simsek and Veiga (2001) suggested, for example, when electronic invitations to participate in a study are sent to employees' e-mail accounts (as we did in the current study), a likely response is, "If they know my e-mail address, they also know me and how I am responding" (p. 231).

A second possibility as to why the latent means from the Web based as compared to the paper-and-pencil application were generally higher involves the type of statistical tests used in our data analysis. Whereas earlier studies (e.g., Booth-Kewley et al., 1992; Eaton & Struthers, 2002; King & Miles, 1995; Knapp & Kirk, 2003; Stanton, 1998) have typically employed ANOVA or simple *t* tests when conducting mean comparisons, we tested for latent-mean differences using mean and covariance structure (commonly referred to as MACS) analysis techniques. As a result, we were able to take into account measurement equivalence, as well as any measurement error. To explore this alternative more fully, we pooled the responses of both the paper-and-pencil and Web-based samples and tested for mean differences across applications using post hoc one-way ANOVAs. Results indicated collective efficacy and collective goal commitment were no longer significantly different across administration modes. This finding supports earlier arguments regarding the advantages of covariance analyses; had we not used MACS when testing for latent-mean differences, we would have erroneously concluded there were no differences between administration modes on two of four study measures. With the availability of numerous statistical packages offering MGCFA and latent-means analysis options (and texts explaining how to use them), we recommend that researchers no longer use traditional ANOVA or *t* tests for making such cross-group comparisons.

A third possibility as to why the means from the Web-based as compared to the paper-and-pencil application were higher concerns the nonrandom assignment of respondents across administration modes. The nonmanagerial nature of the paper-and-pencil sample might, for example, account for the observed lower latent means. We acknowledge this potential limitation but note that in a real-world context, as in the present study, organizations typically do not have the luxury of randomly assigning employees across experimental conditions. More typically, organizations strive for the largest and most representative employee sample possible. This may include administering both paper-and-pencil and Web-based surveys. As a further attempt to address this concern, we conducted a post hoc latent-means comparison test that compared a nonmanagerial subsample ($n = 559$; selected at random from the Web-based sample) and the paper-and-pencil sample of nonmanagerial employees ($n = 559$). The results were entirely consistent with the findings we report for our complete samples.

A second study limitation is likewise related to the nonrandom assignment of respondents. More specifically, the sample sizes across respondent groups differed greatly, possibly influencing parameter standard errors, which could bias nested model comparisons. We

again note that the purpose of the present study was to investigate measurement equivalence issues using data collected in a real-world context. This limitation remains, and therefore, we again conducted post hoc equivalence analyses that included identical sample sizes for both administration modes. Supporting the results based on the complete samples, each measure was found to be fully equivalent across the Web-based and paper-and-pencil administrations.[3]

With few exceptions (e.g., Booth-Kewley et al., 1992; Donovan et al., 2000), past research studies comparing conventional paper-and-pencil and Web-based applications have used undergraduate student samples or recruited (e.g., links placed on Web pages, listservs, search engines, and e-mail snowballing) Internet users (e.g., Buchanan & Smith, 1999; Epstein, Klinkenberg, Wiley, & McKinley, 2001; Herrero & Meneses, in press; King & Miles, 1995; Knapp & Kirk, 2003; Potosky & Bobko, 1997). Although these types of Internet approaches offer certain advantages (e.g., large sample sizes), they also raise concerns about the representativeness of nonrandom samples and the generalizability of the resulting findings to other settings (Kraut et al., 2004). The present study makes a contribution to the measurement literature given that actual employees were sampled as part of an organization-sponsored survey. It thus extends the generalizability of available knowledge to a real-world (workplace) context. We note, however, that respondents in the present study were guaranteed complete anonymity. An alternative survey methodology would have been to assure respondent confidentiality. The difference between these alternatives is that in confidential surveys, respondents can be identified. Consequently, confidential surveys may lead respondents to sanitize their responses for fear of negative evaluations or other forms of reprisal. Additional research is needed, therefore, to determine if administration mode and degree of anonymity interact to influence measurement equivalence.

Although our primary focus has been the psychometric properties of the TL-MLQ across administration modes, demonstrating that the TL-MLQ measure and its associated outcomes were equivalent across a wide variety of nations and languages is notable and, we suggest, may also hold interesting implications for future research and practice. As Scandura noted in her letter exchange with Dorfman, "The need to discern how perceptions of leadership may vary from one culture to another and the implications for understanding seem more pressing than ever in the context of leader emergence today" (Scandura & Dorfman, 2004, p. 280). Accordingly, by our assessment of the TL-MLQ in 13 countries and 12 languages, our results extend international and cross-cultural leadership research (e.g., Den Hartog, House, Hanges, Ruiz-Quintanilla, & Dorfman, 1999), as we have demonstrated that raters of leadership behavior invoked a similar conceptual frame of reference with regard to Bass and Avolio's transformational leadership measure. Furthermore, our results are supportive of Bass's (1997) contention that it may be possible for a single transformational leadership theory to account for leadership across an assortment of differing cultures. Interestingly, previous research that has examined the equivalence of measures oftentimes concludes that the constructs are not equivalent across cultures (e.g., Riordan & Vandenberg, 1994; Ryan et al., 1999). And yet, the present effort adds to the growing body of research (e.g., the GLOBE Project on Leadership Worldwide; Den Hartog et al., 1999) that has demonstrated convergence in cross-national perceptions of leadership. Hence, we echo Scandura's entreaty that the next step will be for researchers to investigate why this so often appears to be the case

(Scandura & Dorfman, 2004). Moreover, leadership researchers are becoming increasingly interested in how leaders exert their influence on followers' job attitudes and behavior. Thus, the analytic approach used in the present study may provide researchers with a powerful tool to investigate the relational equivalence between the TL-MLQ and associated outcomes across cultures and languages. Finally, our results suggest that researchers and practitioners interested in transformational leadership assessment and training and development would benefit from separately assessing lower and higher order transformational leadership constructs. If future research were to find, for example, that the influence of individualized consideration was similar across cultures, such a finding would hold both theoretical and practical implications for multinational corporations in such matters as expatriate training.

There is still much for survey researchers to learn concerning how to design, interpret, and account for responses obtained using multiple modes of administration. The current study was conducted to provide a better understanding of measurement equivalence when "computerizing" a well-developed, paper-and-pencil measure. It is unique in that survey data were collected from respondents in a multinational organization that employed both paper-and-pencil and Web-based surveys to ensure a sample representative of the organization as a whole. Moreover, it extends prior measurement equivalence research on Web-based and paper-and-pencil surveys by exploring the practical question of whether modes of administration affect the strength of the relations among job attitudes and, consequently, how survey results are interpreted by an organization. In sum, our study demonstrates that MGCFA is a useful tool for testing measurement equivalence across research samples and administration modes. Few researchers have explored differences between Web-administered and traditional paper-and-pencil responses collected from actual employees voluntarily participating in an organization-sponsored field study. Although our research incorporated only four measures, our results are consistent with past investigations that have reported few differences between participants' response patterns (e.g., Donovan et al., 2000; Stanton, 1998). Whereas caution is always advised when using multiple modes of administration, our results suggest that, in general, there are minimal measurement differences for well-developed, psychometrically sound instruments whether administered in a paper-and-pencil or online format. In particular, our findings also open a methodological door for survey researchers wishing to assess transformational leadership with a Web-based platform, as well as to compare or combine responses collected with paper-and-pencil and Web-based applications.

Notes

1. Alternative options for dealing with the offending estimate include removing items (Lawley & Maxwell, 1971) and setting the residual variance to zero or some other small value (Chen et al., 2001). We chose not to make use of these alternatives for several reasons. We believed deletion of items to be inappropriate given that we would need to delete items in both samples, even though the offending estimate was found in only the Web-based sample. The second alternative, that is, constraining the negative variance to zero, was also deemed to be unsuitable because this option implies the additional constraint is imposed only on the offending estimate in the Web-based sample. Thus, the baseline models used in subsequent equivalence tests would have been slightly different. Furthermore, the CRDIFF analysis indicated there were instances in both samples involving equal residual variances that, as Byrne (2001) has suggested, are prime candidates for equality constraints. Despite our application of Byrne's recommendation, the equivalence tests were conducted using both alternatives, and the results were nearly identical. These results can be obtained by contacting the first author.

2. The ensuing discussion on the Δcomparative fit index's sensitivity was primarily guided by an anonymous referee's comments.

3. The paper-and-pencil sample consisted of the 665 respondents used in the previous analyses. Using the select cases at random function in SPSS, 5 random samples of 665 respondents were drawn from the 4,244 employees who competed the measures using the Web-based survey, and equivalence analyses were independently conducted across the paper-and-pencil sample and each of the five Web-based subsamples. In each of the equivalence tests, there was no indication of model degradation when additional equality constraints were added to the baseline models.

References

American Psychological Association. (2002). Ethical principles of psychologists and code of conduct. *American Psychologist*, *57*, 1060-1073.

Antonakis, J., Avolio, B. J., & Sivasubramaniam, N. (2003). Context and leadership: An examination of the nine-factor full-range leadership theory using the multifactor leadership questionnaire. *Leadership Quarterly*, *14*, 261-295.

Arbuckle, J. L. (1996). Full information estimation in the presence of missing data. In G. A. Marcoulides & R. E. Schumaker (Eds.), *Advanced structural equation modeling: Issues and techniques* (pp. 243-277). Mahwah, NJ: Erlbaum.

Arbuckle, J. L., & Wothke, W. (1999). *Amos 4.0 user's guide*. Chicago: Small Waters.

Asci, F. H., Eklund, R. C., Whitehead, J. R., Kirazci, S., & Koca, C. (2005). Use of the CY-PSPP in other cultures: A preliminary investigation of its factorial validity for Turkish children and youth. *Psychology of Sport and Exercise*, *6*, 33-50.

Avolio, B. J., Bass, B. M., & Jung, D. I. (1999). Re-examining the components of transformational and transactional leadership using the Multifactor Leadership Questionnaire. *Journal of Occupational and Organizational Psychology*, *72*, 441-462.

Bass, B. M. (1985). *Leadership and performance beyond expectations*. New York: Free Press.

Bass, B. M. (1997). Does the transactional-transformational leadership paradigm transcend organizational and national boundaries? *American Psychologist*, *52*, 130-139.

Bass, B. M., & Avolio, B. J. (2000). *Multifactor Leadership Questionnaire: Technical report, leader form, rater form, and scoring key for MLQ form 5x-short* (2nd ed). Redwood City, CA: Mindgarden.

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

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

Buchanan, T. (2002). Online assessment: Desirable or dangerous? *Professional Psychology: Research and Practice*, *33*, 148-154.

Buchanan, T., Johnson, J. A., & Goldberg, L. R. (2005). Implementing a five-factor personality inventory for use on the Internet. *European Journal of Psychological Assessment*, *21*, 115-127.

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.

Bycio, P., Hackett, R. D., & Allen, J. S. (1995). Further assessments of Bass's (1985) conceptualization of transactional and transformational leadership. *Journal of Applied Psychology*, *80*, 468-478.

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

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

Carless, S. A. (1998). Assessing the discriminant validity of transformational leadership behavior as measured by the MLQ. *Journal of Occupational and Organizational Psychology*, *71*, 353-358.

Chan, D. (1998). Functional relations among constructs in the same content domain at different levels of analysis: A typology of composition models. *Journal of Applied Psychology*, *83*, 234-246.

Chen, F., Bollen, K. A., Paxton, P., Curran, P. J., & Kirby, J. B. (2001). Improper solutions in structural equation models: Causes, consequences, and strategies. *Sociological Methods and Research*, *29*, 468-508.

Cheung, G. W., & Rensvold, R. B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal of Management*, *25*, 1-27.

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

Cohen, R. J., Swerdlik, M. E., & Phillips, S. M. (1996). *Psychological testing and assessment* (3rd ed.). Mountain View, CA: Mayfield.

Den Hartog, D. N., House, R. J., Hanges, P. J., Ruiz-Quintanilla, S. A., & Dorfman, P. W. (1999). Culture specific and cross-culturally generalizable implicit leadership theories: Are attributes of charisma/transformational leadership universally endorsed? *Leadership Quarterly*, *10*, 219-256.

Den Hartog, D. N., Van Muijen, J. J., & Koopman, P. L. (1997). Transactional versus transformational leadership: An analysis of the MLQ. *Journal of Occupational and Organizational Psychology*, *70*, 19-34.

Dillon, W. R., Kumar, A., & Mulani, N. (1987). Offending estimates in covariance structure analysis: Comments on the causes of and solutions to Heywood cases. *Psychological Bulletin*, *101*, 126-135.

Donovan, M. A., Drasgow, F., & Probst, T. M. (2000). Does computerizing paper-and-pencil job attitude scales make a difference? New IRT analyses offer insight. *Journal of Applied Psychology*, *85*, 305-313.

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

Eaton, J., & Struthers, C. W. (2002). Using the internet for organizational research: A study of cynicism in the workplace. *CyberPsychology and Behavior*, *5*, 305-313.

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

Evan, W. M., & Miller, J. R. (1969). Differential effects on response bias of computer vs. conventional administration of a social science questionnaire: An exploratory methodological experiment. *Behavioral Science*, *14*, 216-227.

Fouladi, R. T., McCarthy, C. J., & Moller, N. P. (2002). Paper-and-pencil or online? Evaluating mode effects on measures of emotional functioning and attachment. *Assessment*, *9*, 204-215.

French, D. J., & Tait, R. J. (2004). Measurement invariance in the General Health Questionnaire-12 in young Australian adolescents. *European Child and Adolescent Psychiatry*, *13*, 1-7.

Gosling, S. D., Vazire, S., Srivastava, S., & John, O. P. (2004). Should we trust Web-based studies? A comparative analysis of six preconceptions about Internet questionnaires. *American Psychologist*, *59*, 93-104.

Herrero, J., & Meneses, J. (in press). Short Web-based versions of the Perceived Stress (PSS) and Center for Epidemiological Studies–Depression (CESD) scales: A comparison to pencil-and-paper responses among Internet users. *Computers in Human Behavior*.

Hogg, A. (2002). Conducting online research. *Burk Incorporated White Paper Series*, *3*(2). Retrieved July 8, 2004, from http://www.burke.com/whitepapers/PDF/B.WhitePaperVol3-2002-Iss2.pdf

Hollenbeck, J. R., Klein, H. J., O'Leary, A. M., & Wright, P. M. (1989). Investigation of the construct validity of a self-report measure of goal commitment. *Journal of Applied Psychology*, *74*, 951-956.

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, *3*, 424-453.

King, W. C., Jr., & Miles, E. W. (1995). A quasi-experimental assessment of the effect of computerizing noncognitive paper-and-pencil measurements: A test of measurement equivalence. *Journal of Applied Psychology*, *80*, 643-651.

Knapp, H., & Kirk, S. A. (2003). Using pencil and paper, Internet, and touch-tone phones for self-administered surveys: Does methodology matter? *Computers in Human Behavior*, *19*, 117-134.

Kraut, R., Olson, J., Banaji, M., Bruckman, A., Cohen, J., & Couper, M. (2004). Psychological research online: Report of board of scientific affairs' advisory group on the conduct of research on the Internet. *American Psychologist*, *59*, 105-117.

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

Lawley, D. N., & Maxwell, A. E. (1971). *Factor analysis as a statistical method* (2nd ed.). London: Butterworth.

Lievens. F., & Anseel, F. (2004). Confirmatory factor analysis and invariance of an organizational citizenship behavior measure across samples in a Dutch-speaking context. *Journal of Occupational and Organizational Psychology*, *77*, 299-306.

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

Lowe, K. B., Kroeck, K. G., & Sivasubramaniam, N. (1996). Effectiveness correlates of transformation and transactional leadership: A meta-analytic review of the MLQ literature. *Leadership Quarterly*, *7*, 385-425.

Mantzicopoulos, P., French, B. F., & Maller, S. J. (2004). Factor structure of the pictorial scale of perceived competence and social acceptance with two pre-elementary samples. *Child Development*, *75*, 1214-1228.

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin*, *103*, 391-410.

Meade, A. W., & Lautenschlager, G. J. (2004a). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods*, *7*, 361-388.

Meade, A. W., & Lautenschlager, G. J. (2004b). A Monte-Carlo study of confirmatory factor analytic tests of measurement equivalence/invariance. *Structural Equation Modeling*, *11*, 60-72.

Millsap, R. E., & & Kwok, O. (2004). Evaluating the impact of partial factorial invariance on selection in two populations. *Psychological Methods*, 9, 93-115.

Mullen, M. R. (1995). Diagnosing measurement equivalence in cross-national research. *Journal of International Business Studies*, *26*, 573-596.

Naglieri, J. A., Drasgow, F., Schmit, M., et al. (2004). Psychological testing on the Internet: New problems, old issues. *American Psychologist*, *59*, 150-162.

Ployhart, R. E., & Oswald, F. L. (2004). Applications of mean and covariance structure analysis: Integrating correlational and experimental approaches. *Organizational Research Methods*, *7*, 27-65.

Ployhart, R. E., Wiechmann, D., Schmitt, N., Sacco, J. M., & Rogg, K. (2003). The cross-cultural equivalence of job performance ratings. *Human Performance*, *16*, 49-79.

Potosky, D., & Bobko, P. (1997). Computer versus paper-and-pencil administration mode and response distortion in noncognitive selection tests. *Journal of Applied Psychology*, *82*, 293-299.

Richman, W. L., Kiesler, S., Weisband, S., & Drasgow, F. (1999). A meta-analytic study of social desirability distortion in computer-administered questionnaires, traditional questionnaires, and interviews. *Journal of Applied Psychology*, *84*, 754-775.

Riggs, M. L., & Knight, P. A. (1994). The impact of perceived group success-failure on motivational beliefs and attitudes: A causal study. *Journal of Applied Psychology*, *79*, 755-766.

Riordan, C. M., & 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.

Riordan, C. M., & Weatherly, E. W. (1999). Defining and measuring employees' identification with their work groups. *Educational and Psychological Measurement*, *59*, 310-324.

Rosenfeld, P., Booth-Kewley, S., Edwards, J. E., & Thomas, M. D. (1996). Responses on computer surveys: Impression management, social desirability, and the Big Brother syndrome. *Computers in Human Behavior*, *12*, 263-274.

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

Scandura, T., & Dorfman, P. (2004). Leadership research in an international and cross-cultural context. *Leadership Quarterly*, *15*, 277-307.

Simsek, Z., & Veiga, J. F. (2001). A primer on Internet organizational surveys. *Organizational Research Methods*, *4*, 218-235.

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.

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

Synodinos, N. E., & Brennan, J. M. (1990). Evaluating microcomputer interactive survey software. *Journal of Business and Psychology*, *4*, 483-493.

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.

Whitener, E. M., & Klein, H. J. (1995). Equivalence of computerized and traditional research methods: The roles of scanning, social environment, and social desirability. *Computers in Human Behavior*, *11*, 65-75.

Wothke, W. (2000). Longitudinal and multigroup modeling with missing data. In T. D. Little, K. U. Schnable, & J. Baumert (Eds.), *Modeling longitudinal and multilevel data* (pp. 219-240). Mahwah, NJ: Erlbaum.

Zaccaro, S. J., Rittman, A. L., & Marks, M. A. (2001). Team leadership. *Leadership Quarterly*, *12*, 451-483.

Michael S. Cole is a senior research fellow and lecturer in the Institute for Leadership and Human Resource Management at the University of St. Gallen in Switzerland. His research interests include exploring how organizational contextual factors influence employees' attachments to organizations.

Arthur G. Bedeian is a Boyd Professor at Louisiana State University. He is a former editor of the *Journal of Management,* a past president of the Academy of Management, and a quondam dean of the Academy's Fellows Group.

Hubert S. Feild is Torchmark Professor of Management in the College of Business at Auburn University. His professional interests include human resource selection and research methods in human resource management.