A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research

Robert J. Vandenberg and Charles E. Lance

Organizational Research Methods 2000 3: 4
DOI: 10.1177/109442810031002

A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research

ROBERT J. VANDENBERG
CHARLES E. LANCE
University of Georgia

The establishment of measurement invariance across groups is a logical prerequisite to conducting substantive cross-group comparisons (e.g., tests of group mean differences, invariance of structural parameter estimates), but measurement invariance is rarely tested in organizational research. In this article, the authors (a) elaborate the importance of conducting tests of measurement invariance across groups, (b) review recommended practices for conducting tests of measurement invariance, (c) review applications of measurement invariance tests in substantive applications, (d) discuss issues involved in tests of various aspects of measurement invariance, (e) present an empirical example of the analysis of longitudinal measurement invariance, and (f) propose an integrative paradigm for conducting sequences of measurement invariance tests.

Measurement can be defined as the systematic assignment of numbers on variables[1] to represent characteristics of persons, objects, or events. In the organizational sciences, measurement processes typically are aimed at describing characteristics of individuals, groups, or organizations that are of some substantive interest to the researcher. Measurement has historically been and continues to be a pivotal issue in research in the organizational sciences. Evidence of this includes (a) the number of recent articles that review, evaluate, and recommend measurement practices in the organizational sci-

Authors' Note: Charles E. Lance was supported in part by Grants AG15321-02, National Institutes of Health, and F49620-93-C-0063, Air Force Office of Scientific Research. The authors contributed equally to the development of this article. We are indebted to and wish to thank the editor and the two anonymous reviewers for their invaluable suggestions, contributions, and comments. We are also especially thankful to Gordon Cheung, who spent a great deal of time discussing some of the issues presented in this article and helped us tremendously in cutting through some of the complexity inherent in this topic. Address correspondence concerning this article either to Robert J. Vandenberg, Department of Management, University of Georgia, Athens, GA 30602 (rvandenb@terry.uga.edu) or Charles E. Lance, Department of Psychology, University of Georgia, Athens, GA 30602-3013 (clance@arches.uga.edu).

Organizational Research Methods, Vol. 3 No. 1, January 2000 4-70

ences (e.g., Bagozzi & Edwards, 1998; Bagozzi & Heatherton, 1994; Bagozzi & Phillips, 1991; Hinkin, 1995, 1998; Schmidt & Hunter, 1996); (b) the number of scientific journals devoted at least in part to measurement issues (e.g., *Applied Psychological Measurement*, *Educational and Psychological Measurement*, *Journal of Applied Psychology*, *Journal of Educational Psychology*, *Multivariate Behavioral Research*, *Organizational Research Methods*, *Psychological Methods*, and *Structural Equation Modeling*); (c) the amount of space devoted in methods and discussion sections of empirical journal articles to measurement issues; and (d) the importance accorded to appropriate measurement of key variables in empirical articles in publication decisions (e.g., Campion, 1993). In short, the measurement process is pivotal because it defines the links (through "epistemic definitions") (Cronbach & Meehl, 1955) between organizational theories and the data used to test them.

Historically, evaluation of measurement quality has been rooted in classical test theory (CTT) of true and error scores (Crocker & Algina, 1986; Lord & Novick, 1968; Nunnally & Bernstein, 1994). CTT has provided and will probably continue to provide a solid foundation for the evaluation of manifest (i.e., observed) variables' measurement properties in terms of their reliability and validity. However, additional issues extend beyond the traditional purview of CTT that represent important considerations in evaluating manifest variables' measurement properties, and relatively recent advances in analytic tools have made investigation of these issues much more accessible to researchers. The particular issue that we address here is of measurement equivalence (or, alternately, "measurement invariance") across populations. Example questions underlying measurement equivalence/invariance (ME/I) and that are not directly addressable through traditional CTT avenues are the following: (a) Do respondents from different cultures interpret a given measure in a conceptually similar manner? (b) Do rating sources define performance in similar ways when rating the same target on identical performance dimensions? (c) Are there gender, ethnic, or other individual differences that preclude responding to instruments in similar ways? (d) Does the very process of substantive interest (i.e., an intervention or experimental manipulation) alter the conceptual frame of reference against which a group responds to a measure over time?

The goals of this article are to (a) describe why, although seldom addressed, evaluation of measurement equivalence across populations is a critical issue for organizational researchers; (b) review theoretical/methodological literature on "state-of-the-art" approaches to testing for measurement equivalence; (c) review representative empirical studies that have evaluated measurement equivalence to characterize the state-of-the-practice; (d) present an empirical example of the analysis of longitudinal ME/I; (e) describe an integrated and comprehensive paradigm for conducting evaluations of measurement equivalence; and (f) to point to additional issues that are in need of additional clarification and further research. Our review is confined to evaluation of measurement equivalence in a confirmatory factor-analytic (CFA) framework. As such, we do not review literature relating to other frameworks such as item response theory (e.g., Maurer, Raju, & Collins, 1998; Millsap & Everson, 1993; Reise, Widaman, & Pugh, 1993) and generalizability theory (e.g., Brennan, 1983; Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Marcoulides, 1996, 1998). Also, our discussion is aimed at the evaluation of measurement equivalence of multi-item composite measures, such as climate, job satisfaction, personality, and cognitive ability scales.

The primary theme of this article is to emphasize that a number of specific aspects to the measurement equivalence issue are readily testable within a CFA framework. We feel that it is important to articulate these because (a) knowingly or unknowingly, researchers invoke assumptions about measurement equivalence in conducting tests of substantive hypotheses; (b) although rarely tested, these assumptions are routinely and straightforwardly testable as extensions to the basic CFA framework; and (c) if not tested, violations of measurement equivalence assumptions are as threatening to substantive interpretations as is an inability to demonstrate reliability and validity. To this end, we start with a brief overview of CTT. Within this context, we broaden the CTT measurement model to introduce the ME/I issue and its implications for drawing meaningful substantive inferences from data. Next, we review the existing ME/I literature to identify those tests for ME/I that have been recommended previously and how these tests have been applied in practice. Here, we illustrate the breadth of areas in which measurement equivalence has been investigated and provide a backdrop for an integrative paradigm for conducting tests of measurement equivalence in terms of a recommended sequence of tests.

CTT and Epistemic Definitions in the Measurement Process

One of the most fundamental aspects of CTT (Lord & Novick, 1968) is the definition of observed scores in terms of true and error score components, or

$$X_{ijk} = T_{ij} + E_{ijk}, \tag{1}$$

where X_{ijk} refers to the *k*th observation of the *i*th examinee's score on some measure of the *j*th trait, T_{ij} is the *i*th examinee's true score on trait *j*, and E_{ijk} represents a nonsystematic measurement error score component that, theoretically, is sampled from a normal distribution with zero mean and σ^2_E. T_{ij} is defined as the expected value across realizations of X_{ijk} (i.e., $E[X_{ijk}] = T_{ij}$), and it is assumed that T_{ij} is uncorrelated with E_{ijk} (i.e., $E[T_{ij}, E_{ijk}] = 0$). This latter assumption allows the decomposition of observed score variance (σ^2_X) into true and error score components, or

$$\sigma^2_X = \sigma^2_T + \sigma^2_E, \tag{2}$$

another fundamental definition from CTT that states that variance in individuals' observed scores reflects variance contributed by true individual differences in the measured trait and variance contributed by nonsystematic measurement error. Definition of test reliability in terms of $r_{XX'} = \sigma^2_T / (\sigma^2_T + \sigma^2_E) = \sigma^2_T / \sigma^2_X$ derives directly from Equation (2) and is a major criterion for the evaluation of tests' psychometric properties, which stems from the desire that X_{ijk} should substantially reflect the true score (T_{ij}) to the exclusion of error (E_{ijk}).

Figure 1a shows how these components of CTT and epistemic definitions (theoretical specifications of connections between some latent construct and the operations employed to represent it) (Cronbach & Meehl, 1955) in the measurement process are

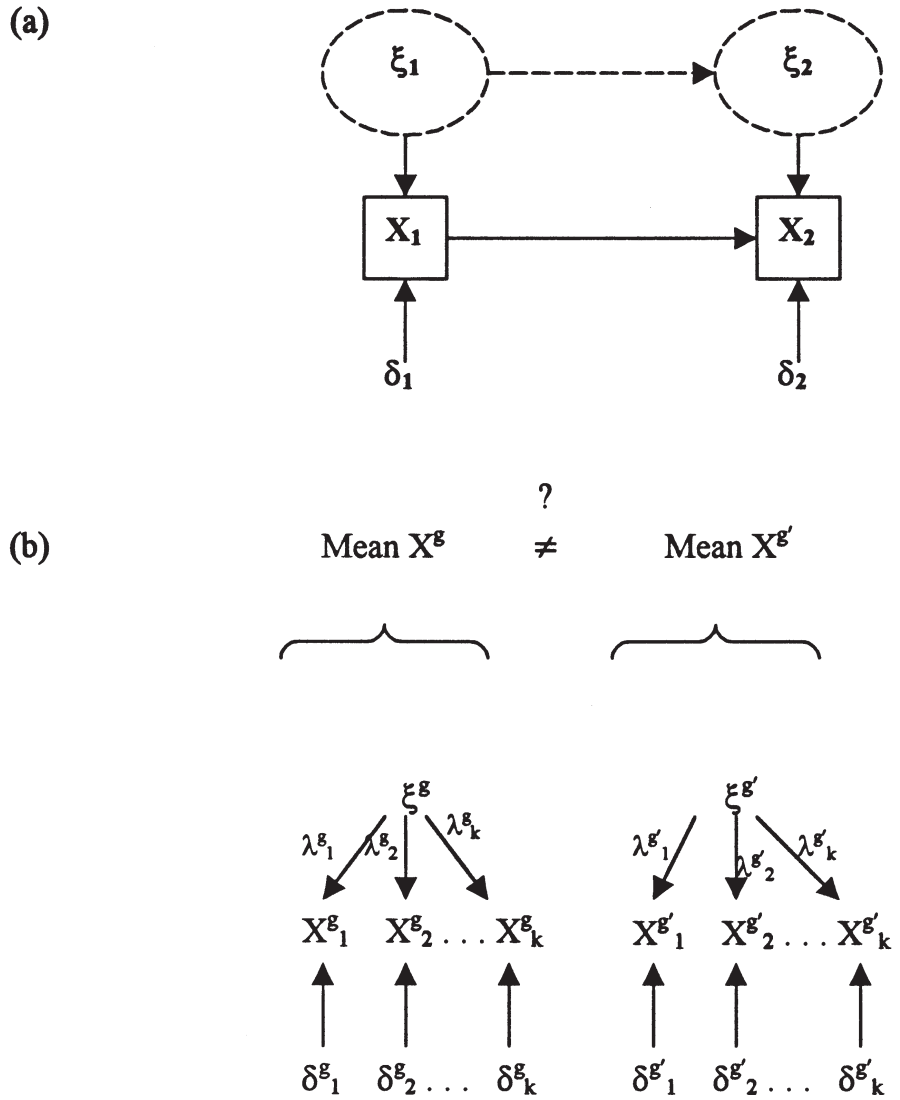

Figure 1: Link Between Observed and True Scores (a) and the Implied Measurement Model Underlying Simple Group Comparisons (b)

closely related. Here, ξ is some hypothetical construct of interest, corresponding to T_{ij}. The dotted circle symbolizes that ξ is an abstraction (latent variable), as is T_{ij}, defined in terms of an expected value of realizations on X_{ijk}. The arrow leading from ξ to X reflects a belief that variation in this abstraction causes variation in X (just as X_{ijk} is assumed to reflect the influence of T_{ij} in Equation (1)). However, the arrow emanating from ξ to X reminds us that measurement systems are imperfect and that some proportion of variation in X may be attributed to influences other than the abstraction of substantive interest, including systematic and nonsystematic measurement error (the

E_{ijk} in Equation (1)). Epistemic definitions connecting X to ξ are of primary concern because it is our confidence in X as a representation of the latent variable, ξ, that supports (approximate) inferences about individuals' standing on ξ and inferences regarding interrelationships between ξ and other theoretical constructs ξ'. It is almost mandatory that authors of empirical studies support their chosen measures' reliability and validity using CTT as the basis of such support (Campion, 1993). Although the latter is unquestionable and most authors adhere to it, our concern is captured in the question, "To what extent are manifest variables' (i.e., *X*s') measurement properties transportable or generalizable across populations?"

Figure 1b provides a simple two-group comparison case to illustrate the importance of this question. Assume as is shown in the top part of Figure 1b that two (or more) groups are administered the measure of interest, some composite score is formed (e.g., an unweighted linear composite of scale items), and the group means on the composite score are compared statistically. For example, a researcher might test the hypothesis that workers in an individualistic culture are less committed than are workers in a collectivist culture (e.g., Riordan & Vandenberg, 1994) by comparing group mean scores on an organizational commitment measure across cultures. Alternately, a researcher might be interested in longitudinal changes in organizational commitment associated with, for example, organizational newcomer socialization practices, organizational interventions, or changes in organizational structure. The analytical approach to the group comparison is unimportant (e.g., a test of mean differences vs. a test of homogeneity of regression, etc.). The key issue here is the desire to make meaningful inferences regarding the status of those groups on the measures and ultimately, perhaps, to draw conclusions as to how this group difference may affect organizational functioning. Applying CTT, authors typically make statements concerning their measure's reliability (stability and/or internal consistency) and validity before undertaking the test of substantive interest. For example, reliability evidence might include a combination of previously reported coefficients with coefficients computed within the current sample. Validity evidence might include (a) evidence of criterion-related validity, noting its application in other studies in which differences occurred or where its hypothesized association held with other measures; (b) evidence of factorial validity, by factor analysis of scale items; (c) discriminate validity evidence by referring to correlations with other measures in the study; or (d) multitrait-mulitmethod construct validity evidence. The main purpose of this reliability and validity evidence is to support the belief that the manifest measure X reflects, as expected, the underlying latent attribute ξ (Nunnally & Bernstein, 1994), that is, that X is a construct-valid operationalization of ξ (Bagozzi & Edwards, 1998). Group comparisons on mean X scores are justified to the extent that psychometric evidence lends confidence that these comparisons approximate comparisons between mean scores on attribute ξ, the theoretical true score.

However, as indicated in the bottom part (below the "{") of Figure 1b, the measurement model underlying the use of observed means (or other composites) is more complex than is often realized. First, Figure 1b indicates that ξ^g implies a content domain in the *g*th group that may only be adequately represented using multiple operationalizations (Embretson, 1983). Second, it indicates that responses on the *k*th operationalization (e.g., the *k*th item in a *K*-item scale) in the *g*th group (X_k^g) reflect, in part, the

intended theoretical construct (this connection is indicated as δ_k^g) and other unique but as yet unaccounted for factors (indicated as δ_k^g), including but not limited to nonsystematic measurement error. Reliability and validity evidence are important for evaluating a measure's general psychometric attributes. However, Figure 1b shows that unambiguous interpretation of observed mean differences is also dependent on the between-group equivalence of the underlying measurement model. Specifically, a comparison between groups as is usually undertaken (i.e., *t* test, ANOVA) and with favorable reliability and validity of the measures still implicitly requires untested assumptions regarding the following:

- conceptual equivalence of the underlying theoretical variable (the latent ξ) in each group,
- equivalent associations (λ_k^g) between operationalizations (X_k^gs) and ξ across groups, and
- the extent to which the X_k^gs are influenced to the same degree and perhaps by the same unique factors (δ_k^g) across groups.

These assumptions are rarely tested, but for reasons articulated in the next section, violation of them can render interpretations of between-group comparisons on the nonequivalent measures highly suspect (Bollen, 1989; Drasgow, 1984, 1987; Vandenberg & Self, 1993). For example, the first assumption (equivalent true scores or theoretical constructs) presupposes that the set of manifest measures evokes the same conceptual frame of reference in each of the comparison groups. However, if one set of measures means one thing to one group and something different to the another group, a group mean comparison may be tantamount to comparing apples and spark plugs. Similarly, if associations (λ_k^g) between like items and the latent variable (ξ) differ across comparison groups, inferences in regard to ξ are compromised because the measures are calibrated to the ξ differently (e.g., a response of "3" on an item may mean something quite different across groups). The crux is that cross-group comparisons require prerequisite assumptions of invariant measurement operations across the groups being compared. As succinctly stated by Horn and McArdle (1992),

> The general question of invariance of measurement is one of whether or not, under different conditions of observing and studying phenomena, measurements yield measures of the same attributes. If there is no evidence indicating presence or absence of measurement invariance—the usual case—or there is evidence that such invariance does not obtain, then the basis for drawing scientific inference is severely lacking: findings of differences between individuals and groups cannot be unambiguously interpreted. (p. 117)

Thus, demonstration of measurement equivalence is a logical prerequisite to the evaluation of substantive hypotheses regarding group differences, regardless of whether the comparison is as simple as a between-group mean differences test or as complex as testing whether some theoretical structural model is invariant across groups. This is most clearly seen by pointing to a number of testable hypotheses regarding measurement equivalence.

Testable Hypotheses Regarding Measurement Equivalence

We assume that some comparison between groups is of importance, and the presence or absence of group differences has some meaningful substantive implications. We assume further that (a) the measure of interest is perceptually based (e.g., requests respondents' reported levels of job satisfaction, organizational commitment, group cohesiveness, etc.); (b) the measure comprises multiple manifest indicators (e.g., multiple scale items); (c) the items are combined additively to operationalize the underlying construct; (d) evidence exists of the measure's psychometric soundness beyond the preliminary stages of scale development; (e) the concern is with effect indicators and not causal indicators (see Bollen & Lenox, 1991); (f) the common factor model holds for describing relationships among items; and (g) the term *groups* refers either to independent groups or to the same group measured longitudinally on multiple occasions. These assumptions are not restrictive because the most organizational research relies on data obtained on measures of the form assumed here (Hinkin, 1998).

Figure 1b is actually a simplification of the complete measurement equivalence issue and represents only a subset of the measurement attributes for which cross-group measurement equivalence must be assumed when conducting group comparisons. To see this, the measurement relationships between the *k* items in the *g*th group in Figure 1b may be represented as

$$\mathbf{X}_k^g = \tau_k^g + \Lambda_k^g \xi^g + \delta_k^g, \tag{3}$$

where $\mathbf{X}_k^g$ refers to the vector of items comprising the composite measure, Λ_k^g refers to the matrix of regression slopes relating the $\mathbf{X}_k^g$ to the ξ^g, τ_k^g refers to the vector of regression intercepts, and δ_k^g refers to the vector of unique factors. Assuming that $E(\xi^g, \delta_k^g) = 0$, the covariance equation that follows from Equation (1) is

$$\Sigma^g = \Lambda_X^g \Phi^g \Lambda_X^{g'} + \Theta_\delta^g, \tag{4}$$

where Σ^g is the matrix of variances and covariances among the *K* items in the *g*th population (group), Λ_X^g is the matrix of items' factor loadings on ξ^g, Φ^g contains variances and covariances among the ξ^g, and Θ_δ^g is the (typically) diagonal matrix of unique variances. Equation (4) is the fundamental covariance equation for factor analysis that models observed item covariances as a function of common (ξ^g) and unique (δ_k^g) factors (see Jöreskog & Sörbom, 1996, p. 3). The inclusion of the τ_k^g vector in Equation (3) is an extension of the basic model only in the sense that in most applications of covariance structure analysis, the intercepts are assumed to be zero and thus are not estimated (see Jöreskog & Sörbom, 1996, p. 297). Thus, the first point of Equations (3) and (4) is to show that aspects of measurement relations can be framed within the common factor model.

The second point of Equations (3) and (4) is to demonstrate the number of specific aspects to the measurement equivalence issue that are testable within a CFA framework. We feel that it is important to articulate these because (a) knowingly or unknow-

ingly, researchers invoke assumptions about measurement equivalence in conducting tests of group differences, and (b) although rarely tested, these assumptions are routinely testable within the general CFA model. As implied from the Horn and McArdle (1992) quote above, not undertaking these tests may have serious implications for the unambiguous interpretation of substantive findings. Specifically, Equations (3) and (4) imply the following as testable hypotheses relating to measurement equivalence:

1. $\xi^g = \xi^{g'}$, that is, that the set of K items evokes the same conceptual framework in defining the latent construct (ξ) in each comparison group (from Equation (3));
2. $\Lambda_k^g = \Lambda_k^{g'}$, that is, that the regression slopes linking the manifest measures X_k^g to the underlying construct(s) are invariant across groups (Equation (3));
3. $\tau_k^g = \tau_k^{g'}$, that is, that the regression intercepts linking the manifest measures X_k^g to the underlying construct(s) are invariant across groups (Equation (3));
4. that the CFA model holds equivalently and assumes a common form across groups (Equation (4));
5. $\Theta_{\delta k}^g = \Theta_{\delta k}^{g'}$, that is, that unique variances for like X_k^gs are invariant across groups (Equation (4)); and
6. $\Phi^g = \Phi^{g'}$, that is, that variances and covariances among the latent variables are invariant across groups (Equation (4)).

Although these aspects of measurement equivalence are testable within a CFA framework, they are rarely evaluated in the organizational literature. We next present a review of literature published in a number of fields that has discussed various tests of measurement invariance (including those above) and their rationale, sequencing, interpretation, and implications for tests of substantive hypotheses. The review is subdivided into two sections—articles focusing on recommended practices and those on applications of the tests. Subsequently, we turn to our recommended paradigm for evaluating measurement equivalence in organizational research.

Review of Measurement Equivalence Literature

We intended our review to be representative of the measurement equivalence/invariance (ME/I) literature but not necessarily exhaustive. To start, we manually reviewed the reference lists of articles of which we were already aware that had conducted tests of ME/I to identify sources that were cited as providing the theoretical rationale for the selected ME/I tests. Second, we compiled a list of the key references that were often cited as providing the theoretical rationale for tests of ME/I (Byrne, 1989; Byrne, Shavelson, & Muthén, 1989; Cole & Maxwell, 1985; Drasgow & Kanfer, 1985; Horn & McArdle, 1992; Jöreskog, 1974; Millsap & Everson, 1991; Millsap & Hartog, 1988; Reise et al., 1993; Rock, Werts, & Flaugher, 1978; Schmitt, 1982) and used the Social Sciences Citation Index (SSCI) and other bibliographic search procedures to identify citations to these references. Our goal was to identify publications that described extensions to the general CFA aimed at testing aspects of ME/I. Our literature review identified several issues that were related but not central to the issue of ME/I (the focus of this article). We excluded these issues from consideration here, not because they are unimportant but rather because of the complexity they add to an already complex issue. Examples of these issues include bodies of literature relating to (a) the theoretical and mathematical background of factorial invariance in exploratory factor analysis or in CFA in which the focus was not on ME/I issues per se (e.g.,

Meredith, 1964a, 1964b, 1993; Mulaik, 1972; Thurstone, 1947); (b) approaches to examining alpha, beta, and gamma change other than CFA-based approaches (e.g., Armenakis & Zmud, 1979; Golembiewski, Billingsley, & Yeager, 1976; Schmitt, Pulakos, & Lieblein, 1984; Terborg, Howard, & Maxwell, 1980); (c) other approaches to change measurement such as latent growth modeling (e.g., Duncan & Duncan, 1995; Duncan, Duncan, & Stoolmiller, 1994; Lance, Meade, & Williamson, in press; McArdle & Anderson, 1990); and (d) general issues regarding CFA and structural equation modeling that did not concern ME/I (e.g., Bollen, 1989; James, Mulaik, & Brett, 1982; Marcoulides & Schumacker, 1996).

Our review indicated that interest in examining ME/I issues from a CFA perspective could be traced to several historical roots, including (a) using CFA to further explicate issues concerning factorial invariance (e.g., Alwin & Jackson, 1981; Horn & McArdle, 1992; Jöreskog, 1974; Marsh, 1994), (b) models for test bias (e.g., Drasgow & Kanfer, 1985; Reise et al., 1993), (c) assessment of longitudinal change (e.g., Schaie & Hertzog, 1985; Schmitt et al., 1984), (d) cross-cultural comparisons (see, e.g., Steenkamp & Baumgartner, 1998, for a comprehensive review), and (e) psychometric scale development (see, e.g., Bagozzi & Edwards, 1998, for a comprehensive review). We also found that ME/I issues were of interest to researchers in a variety of research areas, including gerontology, education, individual differences, cross-cultural psychology, developmental psychology, marketing, criminology, sport psychology, and the organizational sciences. The point here is that issues regarding ME/I are neither new nor poorly understood. ME/I has been examined in many disciplines, under many pretexts, but generally for the same reason—the establishment of ME/I is a precondition for conducting substantive group comparisons.

Summary of recommended practices. Table 1 summarizes recommendations made by authors in articles in which the primary intent was to provide theoretical rationale and describe procedures for testing ME/I in a CFA framework. Next to the source's citation is a statement of the article's main focus and the role of ME/I within that focus. The specific ME/I tests mentioned in the papers are listed across the top of Table 1 according to a common nomenclature that we invoked for the literature review, along with an indication of what aspects of the CFA model were evaluated in testing for ME/I. Generally speaking, ME/I tests were recommended using multisample applications of CFA. In the order that they are reported in Table 1, these are the following:

(a) An omnibus test of the equality of covariance matrices across groups, that is, a test of the null hypothesis of invariant covariance matrices (i.e., $\Sigma^{g} = \Sigma^{g'}$), where g and g' indicate different groups.
(b) A test of "configural invariance," that is, a test of a "weak factorial invariance" null hypothesis (Horn & McArdle, 1992) in which the same pattern of fixed and free factor loadings is specified for each group. Configural invariance must be established in order for subsequent tests to be meaningful.
(c) A test of "metric invariance" (Horn & McArdle, 1992) or a test of a strong factorial invariance null hypothesis that factor loadings for like items are invariant across groups (i.e., $\Lambda^{g} = \Lambda^{g'}$). (At least partial) metric invariance must be established in order for subsequent tests to be meaningful.
(d) A test of "scalar invariance" (Meredith, 1993; Steenkamp & Baumgartner, 1998) or a test of the null hypothesis that intercepts of like items' regressions on the latent variable(s) are invariant across groups (i.e., $\tau^{g} = \tau^{g'}$).

(e) A test of the null hypothesis that like items' unique variances are invariant across groups (i.e., $\Theta^{g} = \Theta^{g'}$). Tests (c) through (e) follow the same sequence as that recommended by Gulliksen and Wilks (1950) for tests of homogeneity of regression models across groups and should be regarded as being similarly sequential, so that tests of scalar invariance should be conducted only if (at least partial) metric invariance is established, and tests of invariant uniquenesses should proceed only if (at least partial) metric and scalar invariance has been established first.

(f) A test of the null hypothesis that factor variances were invariant across groups (i.e., $\Phi_{j}^{g} = \Phi_{j}^{g'}$). This was sometimes treated as a complement to Test (c), in which differences in factor variances were interpreted as reflecting group differences in the calibration of true scores (e.g., Schaubroeck & Green, 1989; Schmitt, 1982; Vandenberg & Self, 1993).

(g) A test of the null hypothesis that factor covariances were invariant across groups (i.e., $\Phi_{jj'}^{g} = \Phi_{jj'}^{g'}$). This was sometimes treated as a complement to Test (b), in which differences in factor covariances were interpreted as reflecting differences in conceptual associations among the true scores (e.g., Schmitt, 1982). Often, this and Test (f) were combined in an omnibus test of the equality of the latent variables' variance/covariance matrices across groups (i.e., $\Phi^{g} = \Phi^{g'}$).

(h) A test of the null hypothesis of invariant factor means across groups (i.e., $\kappa^{g} = \kappa^{g'}$), which often was invoked as a way to test for differences between groups in level on the construct of interest.

(i) Other, more specific tests that were discussed by only one or a few of the sources reviewed.

We maintain Byrne et al.'s (1989) distinction in Table 1 by referring to the first five of these tests as tests of aspects of measurement invariance (as they concern tests of relationships between measured variables and latent constructs) versus the next three as testing aspects of structural invariance (as they refer to tests concerning the latent variables themselves). As indicated in the body of Table 1, if an explicit order or hierarchy of tests was recommended, this order was denoted as "Step 1," "Step 2," and so forth, so that the test labeled "Step 1" occurred before "Step 2" and subsequent steps. If a particular step appears more than once, this indicates that the tests labeled in the same step appeared to be conducted simultaneously. Second (when possible), we report the name accorded for each test by the sources' authors, where quotation marks indicate at least close paraphrases for the label. Our goal was to examine consistencies and inconsistencies across published sources with respect to recommendations as to (a) how to test various aspects of ME/I, (b) which tests should be undertaken, (c) the sequencing or order in which the various tests should be evaluated, and (d) inferences accorded to each test endorsed.

The greatest consistency appeared in how various tests should be effected. Most authors were faithful to Jöreskog's (1971) original recommendations as to the application of multisample covariance structure analysis. Thus, for example, when advocating a test of configural invariance, authors did so by specifying the same pattern of fixed and free elements in the factor pattern matrix for each group. Similarly, tests of metric invariance were (appropriately) effected by constraining factor loadings to be invariant (equal) across groups. Furthermore, null hypotheses of equivalence were most often tested by evaluating the fit of some restricted model (e.g., $\Lambda^{g} = \Lambda^{g'}$) relative to some less restrictive baseline model (often the model that specified only configural invariance). However, even here inconsistencies emerged. For example, Horn and McArdle (1992) used a "step-down" approach in which the baseline model was one in

Table 1
Summary of Recommended Practices

			Measurement Invariance (MI)					*Structural Invariance (SI)*			
Reference	*Primary Issue*	*Specific MI Issue*	*Invariant Covariance* ($\Sigma^g = \Sigma^{g'}$)	*Configural Invariance*	*Metric Invariance* ($\Lambda^g = \Lambda^{g'}$)	*Scalar Invariance* ($\tau^g = \tau^{g'}$)	*Invariant Uniquenesses* ($\Theta^g = \Theta^{g'}$)	*Invariant Factor Variances* ($\phi_j^g = \phi_j^{g'}$)	*Invariant Factor Covariances* ($\phi_{jj'}^g = \phi_{jj'}^{g'}$)	*Equal Factor Means* ($\kappa^g = \kappa^{g'}$)	*Other Tests*
Alwin and Jackson (1981)	Factorial invariance Across populations	Confirmatory factor analysis (CFA) as alternative to exploratory factor analysis (EFA)-based assessments of factorial invariance	Step 1a: Omnibus test of MI	(Assumed)	Step 2a: Invariant factor patterns		Step 3: Invariant disturbance covariance structures	Step 4: Invariant factor variance and covariance structures		Step 1b: If $\Sigma^g = \Sigma^{g'}$ Step 2b: If $\Lambda^g = \Lambda^{g'}$	
Bagozzi and Edwards (1998)	Construct validation	Test of MI (generalizability of measurement) as one aspect of construct validation	Step 1: Equality of covariance matrices	Step 2: Baseline (equality of factor patterns)	Step 3: Factor loadings invariant		Step 4: Error variances invariant	Step 5: Invariant factor correlations		Step 6: Equality of factor means	Tests of partial invariance at Step 3
Byrne, Shavelson, and Muthén (1989)	Measurement invariance	Partial MI across populations	Step 1: Equivalent Covariance matrices	Step 2: Equivalent factor structures	Step 3: Invariant factor loadings			Step 4: Invariant structural parameters		Step 5: Invariance of mean structures	Step 3a: Some Λ^g = some $\Lambda^{g'}$, partial invariance Step 4a: Some ϕ^g = Some $\phi^{g'}$, Partial Structural invariance
Chan (1998)	Latent growth Modeling (LGM)	MI as a logical and analytical prerequisite to LGM	Step 1: Invariant factor patterns	Step 2: Invariant factor loadings				Step 3: Invariant factor variances and covariances		Step 4: Invariant factor means	LGM of longitudinal relationships

Cole and Maxwell (1985)	Analysis of MTMM data	Functional equivalence of traits across populations	Step 1: Test of Var/COV equality	Step 2: Equality of factor structures	Step 3: Equality of scaling units	Step 4b: If $\phi_j^g = \phi_j^{g'}$ Step 7: $\phi_j^g / (\phi_j^g + \theta_{jk}^g) = \phi_j^{g'} / (\phi_j^{g'} + \theta_{jk}^{g'})$ If $\phi_j^g \neq \phi_j^{g'}$	Step 4a: Equality of factor variances	Step 6: Equality of constructs across populations		Step 5: $\phi^g{}_{traits,\ methods} = \phi^{g'}{}_{traits,\ methods} = 0$
Drasgow and Kanfer (1985)	Measurement equivalence of attitude scales across populations			Step 1: Equivalent factor structures	Step 2: Metric comparability	Step 2a: Heterogeneous uniqueness				
Horn and McArdle (1992)	Measurement invariance	Metric invariance	Implied by (Step 1)	Compared to metric Invariance, Considered as a weaker form of Factorial Invariance	Step 1: Plus $\phi_j^g = \phi_j^g$ plus $\phi_{jj'}^g = \phi_{jj'}^{g'}$ plus $\kappa^g = \kappa^{g'}$ plus $\kappa^g = \kappa^{g'}$ Plus $\phi_j^g = \phi_j^{g'}$ plus $\kappa^g = \kappa^{g'}$ Step 3: Factorial Invariance		Step 1: Plus $\Lambda^g = \Lambda^{g'}$ plus $\phi_{jj'}^g = \phi_{jj'}^{g'}$ plus $\kappa^g = \kappa^{g'}$ Step 2: Plus $\Lambda^g = \Lambda^{g'}$	Step 1: Plus $\Lambda^g = \Lambda^{g'}$ plus $\phi_j^g = \phi_j^{g'}$ plus $\kappa^g = \kappa^{g'}$	Step 1: Plus $\Lambda^g = \Lambda^{g'}$ plus $\phi_j^g = \phi_j^{g'}$ plus $\phi_{jj'}^g = \phi_{jj'}^{g'}$ plus Step 1: Plus $\Lambda^g = \Lambda^{g'}$ plus $\phi_j = \phi_j^{g'}$ plus	
Jöreskog (1971)	Factorial Invariance Across populations	Application of SIFASP to assess MI	Step 1: Equality of covariance matrices	Step 2: Equality of number of common factors	Step 3: Invariant factor pattern	Step 4: Invariant uniqueness	Step 5 Invariant factor variances and covariances			
Marsh (1994)	Factorial Invariance Across populations	Sequential ordering of tests of MI		Step 1: Baseline model	Step 2: Factor loading invariance	Step 5: Invariant uniqueness	Step 4: Invariant factor variances	Step 3: Invariant factor covariances		Test of interactive effects of sub-population membership on aspects of MI
Reise, Widaman, and Pugh (1993)	CFA versus IRTapproaches to assessing MI	Partial MI		Partial MI Baseline model	Step 2: Full measurement invariance					Step 3: Some Λ^g = some $\Lambda^{g'}$, partial measure-ment invariance

(continued)

Table 1 Continued

			Measurement Invariance (MI)					*Structural Invariance (SI)*			
Reference	*Primary Issue*	*Specific MI Issue*	*Invariant Covariance* ($\Sigma^g = \Sigma^{g'}$)	*Configural Invariance*	*Metric Invariance* ($\Lambda^g = \Lambda^{g'}$)	*Scalar Invariance* ($\tau^g = \tau^{g'}$)	*Invariant Uniquenesses* ($\theta^g = \theta^{g'}$)	*Invariant Factor Variances* ($\phi_j^g = \phi_j^{g'}$)	*Invariant Factor Covariances* ($\phi_{jj'}^g = \phi_{jj'}^{g'}$)	*Equal Factor Means* ($\kappa^g = \kappa^{g'}$)	*Other Tests*
Rock, Werts, and Flaugher (1978)	Homogeneity of covariance matrices for MANOVA	Sources of MI in covariance matrices	Step 1: Test of homogeneity of covariance matrix	Step 2: Equality of factors models across populations	Step 3: Equality of scaling units	Step 6: Equality of intercepts	Step 5a: If $\phi_j^g = \phi_j^{g'}$ Step 5b: $\phi_j^g / (\phi_j^g + \theta_{jk}^g) = \phi_j^{g'} / (\phi_j^{g'} + \theta_{jk}^{g'})$ if $\phi_j^g \neq \phi_j^{g'}$	Step 4: Equality of true variances		Step 7: Equality of factor means	
Schaie and Hertzog (1985)	Psychometrics/ measurement in aging research	Measurement equivalence as an overlooked issue in measurement	Step 1: Equivalent population covariance matrices	Step 2: Equality of factor models	Step 3: Equality of scale units		Step 5: Equality of reliabilities	Step 4: Equality of true score variances			
Steenkamp and Baumgartner (1998)	MI in multinational consumer research	Proposed paradigm for testing for MI	Step 1: Equality of covariance matrices and mean vectors	Step 2: Configural invariance	Step 3: Full metric invariance	Step 4: Full scalar invariance	Step 7: Error variance invariance	Step 6: Factor variance invariance	Step 5: Factor covariance invariance		Partial MI tested at each step
Taris, Bok, and Meijer (1998)	Stability of measures' Psychometric Properties	Proposed paradigm for testing for MI		Step 1: Equal number of factors and factor patterns	Step 2: Equality of magnitude of factor loadings	Step 6: Equalof item means	Step 5: Equality error variances	Step 4: Equality of constructs' variances	Step 3: Equal factor covariances	Step 7: Equality of latent means	Step 8: Normative stability (over-time factor covariances)

which all invariance restrictions of interest were implemented first, and subsequent models relaxed series of invariance restrictions.

In terms of which tests of ME/I should be undertaken, authors' recommendations varied somewhat across sources. For example, no source recommended conducting all tests listed in the first eight columns in Table 1. In addition, omitted tests varied from article to article. For example, whereas Cole and Maxwell (1985) presented the merits of all tests except scalar invariance and equal factor means, Taris, Bok, and Meijer (1998) discussed the latter two tests but omitted reference to an omnibus test of the equality of groups' covariance matrices. This is not to imply that any of these sources are in error or that their treatment of ME/I issues was incomplete, as the focus varied across these articles, but merely to suggest that these papers make somewhat different (although overlapping) recommendations as to which ME/I tests can or should be conducted.

The frequency with which each test was discussed also varied considerably across sources. Metric invariance (invariant factor loadings) was discussed most frequently, followed by tests of configural invariance (equivalent specification of free and fixed patterns of factor loadings) and of invariant uniqueness across groups. Inasmuch as issues of ME/I are rooted historically in the factorial invariance literature, the frequency with which these tests were advocated is not surprising. In contrast, a test of scalar invariance (invariant intercepts) was discussed least frequently, followed by the test of equal factor (latent) means. This also is not surprising because (a) location parameters (intercepts) are often treated as being arbitrary and sample specific, and (b) analysis of covariance and mean structures is a relatively recent development in the structural equation modeling literature. Thus, the infrequency of those tests does not mean to imply that some authors omitted reference to important aspects of ME/I that others did discuss.

Table 1 also shows that there was variability in terms of the recommended sequence of ME/I tests. For example, there was general agreement that an omnibus test of the equality of the covariance matrices should be conducted first (e.g., Alwin & Jackson, 1981; Bagozzi & Edwards, 1998; Byrne et al., 1989; Cole & Maxwell, 1985; Horn & McArdle, 1992; Jöreskog, 1971; Rock et al., 1978; Schaie & Hertzog, 1985; Steenkamp & Baumgartner, 1998). Researchers have agreed that if covariance matrices do not differ across groups, then ME/I is established, and further tests of the other aspects of measurement equivalence are not necessary (e.g., Alwin & Jackson, 1981; Bagozzi & Edwards, 1998; Cole & Maxwell, 1985; Jöreskog, 1971; Mulaik, 1975; Steenkamp & Baumgartner, 1998). If, however, groups' covariance matrices differ, then investigation into the specific source of the lack of equivalence is warranted. Second, there was broad consensus that configural invariance should be tested next (if groups' covariance matrices are found to differ). It should be tested next because configural invariance must be established as a necessary condition for the evaluation of further aspects of ME/I. As such, the configural invariance model was often viewed as a baseline model against which further tests of ME/I are evaluated. It is a necessary condition because if configural invariance is not demonstrated across groups, then further tests are unwarranted because observed measures represent different constructs within each group.

Beyond this point, however, there was much less agreement as to what constituted the proper sequencing of tests. Most authors advocated testing metric invariance after configural invariance had been established as a baseline model, but not all authors

agreed. As noted by Bollen (1989), Marsh (1994), and most recently by Bagozzi and Edwards (1998), lack of a clear consensus as to the ordering or sequencing of tests beyond this general level might be expected. That is, like all statistical analyses, the aims and goals of the study should determine which specific ME/I tests are undertaken and the ordering in which they are undertaken. We agree that there may not be a sequence of ME/I tests that is universally applicable for every application. However, we argue that tests of measurement invariance (associations of observed scores to the latent variable or variables) should precede tests of structural invariance (associations of latent variables with each other). Our logic is based on Anderson and Gerbing's (1988) argument that one needs to understand what one is measuring before testing associations among what is measured.

The rightmost column of Table 1 shows that several articles also discussed other tests related to ME/I issues. Some of these represented the source's unique focus and contribution. For example, Chan (1998) discussed ME/I tests as logical prerequisites to conducting latent growth modeling of longitudinal data, Cole and Maxwell (1985) discussed ME/I tests as they related to multitrait-multimethod analyses, Marsh (1994) discussed interactive effects of group membership on ME/I, and Taris et al. (1998) discussed measurement stability as a separate concern in tests of ME/I. However, the most frequent additional tests were those of partial invariance (Bagozzi & Edwards, 1998; Byrne et al., 1989; Reise et al., 1993; Steenkamp & Baumgartner, 1998). The logic of testing for partial ME/I is that invariance restrictions may hold for some but not all manifest measures across populations, and relaxing invariance constraints where they do not hold controls for partial measurement inequivalence. Even here, however, there were differences in emphasis across sources. For example, Bagozzi and Edwards (1998) and Reise et al. (1993) discussed tests of partial ME/I only in reference to tests of metric invariance, whereas Byrne et al. (1989) discussed partial ME/I in reference to metric and structural invariance, and Steenkamp and Baumgartner (1998) recommended partial ME/I tests at each step in their recommended sequence.

Finally, we note that the nomenclature invoked to describe the various ME/I tests listed in Table 1 varied across sources. For example, the test of configural invariance was referred to as a "baseline" model (Bagozzi & Edwards, 1998; Marsh, 1994; Reise et al., 1993), a test of "equality of factor structures" (Cole & Maxwell, 1985), or of "equal number of factors and factor patterns" (Taris et al., 1998). The test of metric invariance was also referred to variously as a test of "invariant factor patterns" (Alwin & Jackson, 1981), "equality of scaling units" (Cole & Maxwell, 1985; Schaie & Hertzog, 1985), "metric comparability" (Drasgow & Kanfer, 1985), "factorial invariance" (Horn & McArdle, 1992), "factor loading invariance" (Marsh, 1994), and "full measurement invariance" (Reise et al., 1993). Similarly, the test of invariant uniquenesses was referred to as a test of "invariant disturbance covariance structures" (Alwin & Jackson, 1981), "invariant error variances" (Bagozzi & Edwards, 1998; Steenkamp & Baumgartner, 1998; Taris et al., 1998), and "equality of reliabilities" (Schaie & Hertzog, 1985). These terminological differences created difficulty in making linkages between methodological approaches proposed or adopted across different sources. Therefore, we strongly recommend the nomenclature listed at the top of Table 1 as a common one for future studies on ME/I issues because it is technically accurate and neutral with respect to substantive concerns.

To summarize Table 1, although there was general agreement as to the specific mechanics of conducting various tests of ME/I (e.g., constraining factor loadings to be equal across groups to test metric invariance), there was little consensus among the sources as to the set of tests that constitutes a thorough examination of ME/I. Also, there was little consensus on the particular sequence of tests that should be conducted, although there was general agreement that (a) an omnibus test of equality of covariance matrices across groups was an important first step, and (b) a test of configural invariance is necessary and can serve as a baseline model for further tests. Also, although several sources advocated tests of partial ME/I, there was little consensus as to what aspects should be tested or in what sequence. Finally, differences in nomenclature often made comparisons across sources difficult. For this reason, we offer the common nomenclature for the various tests of ME/I listed in Table 1. Given the diversity of substantive disciplines on which our literature review was based, the lack of consensus on these issues was not surprising.

Summary of applied practices. We limited studies for this section primarily to those in which ME/I was an issue within some broader substantive context. Many of the studies thus identified cited the papers listed in Table 1 in connection with some other aspect of the cited paper (e.g., many studies cited the Cole & Maxwell, 1985, paper in connection with multitrait-multimethod analysis) or in reference to ME/I issues in general. Through this process, we identified the 67 studies listed in Table 2. Unlike the papers listed in Table 1 in which a primary theme was developing ME/I as an issue, studies listed in Table 2 conducted ME/I tests in conjunction with and generally subordinate to tests of substantive hypotheses.

Table 2 summarizes aspects of these empirical studies in much the same manner as Table 1. However, Table 2 also segregates studies according to topical subareas. Focusing first on those subareas, it is seen that ME/I issues have been investigated in a variety of contexts. For example, researchers have used ME/I as a means to operationalize alpha, beta, and gamma change within organizational settings (e.g., Schaubroeck & Green, 1989; Schmitt, 1982; Vandenberg & Self, 1993). Other researchers have tested for ME/I to supplement and extend traditional tools in the scale development and validation process (e.g., Burke, Brief, George, Roberson, & Webster, 1989; Manolis, Levin, & Dahlstrom, 1997). Still others have tested for ME/I procedures to examine equivalence of alternative test administration modes (i.e., video based vs. paper and pencil) (Chan & Schmitt, 1997; King & Miles, 1995). Another popular application of ME/I tests was addressing aspects of cross-cultural generalizability of measures and models (e.g., Chan, Schmitt, Sacco, & DeShon, 1998; Little, 1997; Palich, Hom, & Griffeth, 1995; Riordan & Vandenberg, 1994). A major aim of these cross-cultural studies was to determine whether the same measure evoked the same cognitive frame of reference across cultures and, consequently, whether cross-cultural comparisons on these measures can legitimately be undertaken. A similar motivation underlay the studies listed in the "cross-group comparison" subgroups. Included here are tests of ME/I between (a) organizational levels (e.g., Byrne, 1991), (b) ethnic groups (e.g., Tansy & Miller, 1997), (c) gender (e.g., Stacy, MacKinnon, & Pentz, 1993), (d) age groups (e.g., Babcock, Laguna, & Roesch, 1997), and (e) multiple grouping variables simultaneously ("multiple applications/dimensions," e.g., gender and age groups) (e.g., Marsh, 1993; Schulenberg, Shimizu, Vondracek, & Hostetler, 1988). Finally,

(Text continues on page 34)

Table 2
Summary of Applications

			Measurement Invariance (MI)					*Structural Invariance (SI)*			
Reference	*Primary Issue*	*Specific MI Issue*	*Invariant Covariance* ($\Sigma^{g} = \Sigma^{g'}$)	*Configural Invariance*	*Metric Invariance* ($\Lambda^{g} = \Lambda^{g'}$)	*Scalar Invariance* ($\tau^{g} = \tau^{g'}$)	*Invariant Uniquenesses* ($\theta^{g} = \theta^{g'}$)	*Invariant Factor Variances* ($\phi_{j}^{g} = \phi_{j}^{g'}$)	*Invariant Factor Covariances* ($\phi_{jj'}^{g} = \phi_{jj'}^{g'}$)	*Equal Factor Means* ($\kappa^{g} = \kappa^{g'}$)	*Other Tests*
Alpha, Beta, and Gamma (A, B, Γ) changes											
Bartunek and Franzak (1988)	Organizational restructuring's impact on Employees' frames of reference	Test whether gamma changes occur in frames of reference across time		Step 1: Test of gamma change							
Finch, Okun, Barrera, Zautra, and Reich (1989)	Effects of elder disability on psychological well-being	Test MI across controls versus elders with various types of disabilities	Step 2: Equivalent covariance matrices	Step 1: Generalizable factor structure	Step 3: Equivalent factor loading matrix		Step 5: Equivalence of error structure	Step 7: Invariant latent factor variances	Step 6: Equality of covariances		Step 4: Some Λ^{g} = some $\Lambda^{g'}$ (partial measurement invariance′) Step 8: Some ϕ_{j}^{g} = some $\phi_{j}^{g'}$
Millsap and Hartog (1988)	Detection of ABΓ change	Specific procedures to detect ABΓ change in pretest-posttest Designs		Step 1: baseline model	Step 2: Constraint imposed for posttest groups only; test for Differential gamma change				Step 4: Pretest $\kappa_{control} = \kappa_{exp}$; test of pretreatment equivalence		Step 3: Equality of regression slopes' from posttest scores' regressions on pretest scores (i.e., $\Gamma_{control} = \Gamma_{Experimental}$); differential beta Change Step 5: Equality of Regression intercepts from posttest scores' regressions on pretest scores (i.e., $\alpha_{control} = \alpha_{exp}$); invariant

Pentz and Chou (1994)	Assessment of change from Clinical interventions	Explicate techniques to assess cross-group and longitudinal change		(Equivalent measurement models developed from Step 1)	Step 3a: invariance test of factor loadings	Step 2a Equality of intercepts across groups	Step 4a Invariance test of measurement errors	Step 5a: Invariance test of factor correlations			latent residualized change scores Step 1: Prelimi-nary EFA to develop measurement models Step 2b (etc.): Partial invariance examined here and in subsequent steps
Schaubroeck and Green (1989)	Change in work-related Perceptions During Organizational entry	Assessment of ABΓ change	Step 1: Equality of covariance matrixes	Step 2: Assessment of change in factor structure, test of Gamma change	Step 5: Recalibration of true score units; test of beta change			Step 4: Recalibration of true score continua; test of beta change	Step 3: Equivalence of common factor covariances; test of gamma change	Step 6: Equality of factor location parameters; test of alpha change	
Schmitt (1982)	Assessment of change in organizations	Detection of ABΓ change	Step 1: Equality of pretest/ Posttest Covariance matrices	Step 2: Similarity of factor structure; test of gamma change	Step 5 Equality of factor loadings; test of beta change			Step 4: Equality of factor variances; test of beta change	Step 3: Equality of factor covariances; test of gamma change		
Schmitt, Pulakos, and Lieblein (1984)	Comparison of techniques to assess beta and gamma change	One technique to assess beta and gamma change	Step 1: Equality of pretest/ posttest covariance matrices	Step 2: Similarity of factor structure; test of gamma change	Step 4: Equality of factor loadings; test of beta change		Step 5: Test of equal reliabilities	Step 3: Equality of factor variance-covariance matrices; test of gamma change			
Vandenberg and Self (1993)	Newcomer work adjustment	Assessment of ABΓ change in work-related attitudes	Step 1: Equality of covariance matrices	Step 2: First test of gamma change	Step 5: Extended test of beta change			Step 4: First test of beta change	Step 3: Equality of factor covariance; extended test of gamma change	Step 6: Equality of latent means	For all tests, allowed correlated residuals for identical items across measurement occasions

(continued)

Table 2 Continued

			Measurement Invariance (MI)					*Structural Invariance (SI)*			
Reference	*Primary Issue*	*Specific MI Issue*	*Invariant Covariance* ($\Sigma^g = \Sigma^{g'}$)	*Configural Invariance*	*Metric Invariance* ($\Lambda^g = \Lambda^{g'}$)	*Scalar Invariance* ($\tau^g = \tau^{g'}$)	*Invariant Uniquenesses* ($\theta^g = \theta^{g'}$)	*Invariant Factor Variances* ($\phi_j^g = \phi_j^{g'}$)	*Invariant Factor Covariances* ($\phi_{jj'}^g = \phi_{jj'}^{g'}$)	*Equal Factor Means* ($\kappa^g = \kappa^{g'}$)	*Other Tests*‘
Scale development/validation issues											
Babyak, Snyder, and Yoshinobu (1993)	Psychometric properties of the Hope Scale	Generalizability of scale properties across gender		Step 2: Equivalence of (second-order) factor structure	Step 3: Equivalence of first-order factor (FOF) loadings Step 4: Equivalence of second-order factor (SOF) loading		Step 5: Equivalence of FOF unique variance terms				Step 1: Preliminary EFA to develop FOF measurement models
Burke, Brief, George, Roberson, and Webster (1989)	Cross-validation of the job affect scale	Generalizability across samples		Step 1: Invariant number of factors	Step 2: Invariant factor loadings		Step 3: Invariant error/uniqueness		Step 4: Invariant actor covariances		
Byrne (1994)	Cross-validation of the Maslach Burnout inventory	MI across gender			Step 1: Equality of factor loadings and cross-loadings				Step 2: Equality of factor covariances		Step 3: Equality of error covariances
Byrne and Baron (1993)	Cross-validation of the Beck Depression Inventory	MI across three samples of adolescents		Step 2: Equivalence of SOF structure	Step 3a: Equality of FOF loadings Step 4: Equality of SOF loadings						Step 1: Preliminary EFA to develop FOF measurement models Step 3b: Partially invariant FOF loadings
Ferrando (1996)	Demonstration of CFA to Evaluate Measurement scale properties	MI across gender			Step 1: Baseline model	Step 2: Partial invariance	Step 3: Strong factorial invariance	Step 4: Strict factorial invariance			

Finch and West (1997)	Review of statistical methods for the study of Personality structure	MI as one issue	Step 1: Overall test of equivalence of covariance matrices	Step 2: Same factor structure	Step 3: Equal factor loadings	Step 4: Equality of specific factors	Step 5 Equality of factor covariances			
Manolis, Levin, and Dahlstrom (1997)	Development and evaluation of a Generation X Scale	MI to conduct known groups validation		Step 1: Test of form	Step 2: Factor coefficient invariance test	Step 5: Equality of error variances		Step 4: Invariant factor intercorrelations	Step 3: Invariant factor means	
Soeken and Prescott (1991)	Development and evaluation of the Patient Intensity Nursing Index	MI to cross-validate results across samples		Step 1: Factor pattern invariance	Step 2: Factor loadings invariance	Step 3: Invariance of error/covariance matrix				
Alternative test administration modes										
Chan and Schmitt (1997)	Subgroup differences with respect to situational judgment test scores	MI with respect to video versus paper-and-pencil test administration mode		Step 1: Common factor solution across groups	Step 2: Equal factor loadings	Step 3: Equal error variances		Step 4: Equal factor covariances		
Hattrup, Schmitt, and Landis (1992)	Measurement equivalence of alternative cognitive ability tests	MI of job specific versus commercially available tests		Step 2: Congeneric test model	Step 3: Tau-equivalent test model	Step 4: Parallel test model				Step 1: Specification of null baseline model Step 5: Partially invariant uniqenesses; some (θ^{g} = some $\theta^{g'}$) Step 6: Partially invariant factor loadings (some Λ^{g} = some $\Lambda^{g'}$)
Hong (1995)	State versus trait regulation models	MI as test of state versus trait models		Step 1: Simultaneous FOF solution	Step 2: Equality of factor loadings			Step 3: Equality of factor correlations		Step 4: Partial invariance of factor correlations

(continued)

Table 2 Continued

Reference	Primary Issue	Specific MI Issue	Measurement Invariance (MI): Invariant Covariance ($\Sigma^g = \Sigma^{g'}$)	Configural Invariance	Metric Invariance ($\Lambda^g = \Lambda^{g'}$)	Scalar Invariance ($\tau^g = \tau^{g'}$)	Invariant Uniquenesses ($\theta^g = \theta^{g'}$)	Structural Invariance (SI): Invariant Factor Variances ($\phi_j^g = \phi_j^{g'}$)	Invariant Factor Covariances ($\phi_{jj'}^g = \phi_{jj'}^{g'}$)	Equal Factor Means ($\kappa^g = \kappa^{g'}$)	Other Tests‘
Alternative test administration modes											
King and Miles (1995)	Comparisons of test Administration modes for Noncognitive measures	Computerized versus paper-and pencil measures		Step 2: Equal number of factors	Step 3: Equal factor loadings						Step 1: Specification of null baseline model
Schmit, Ryan, Stierwalt, and Powell (1995)	Impact of respondents' frame of reference on their responses to personality stests	MI to evaluate equivalence of alternative instructional sets		Step 1: Equal structure	Step 3: Equal loadings		Step 4: Equal error variances		Step 2: Equal factor correlations		
Van de Vijver and Harsveld (1994)	Comparison of test administration modes for the GATB	Computer-administered versus paper-and-pencil modes	Step 1: Equal covariance matrices	Step 3: Equal factor model	Step 6: Equal factor loadings			Step 5: Equal factor variances	Step 4: Equal factor covariances		Step 2: Specification of null base line model
Cross-cultural comparisons											
Dumka, Stoerzinger, Jackson, and Roosa (1996)	Cross-cultural and linguistic equivalence of the parenting Self-agency measure	Equivalence across Anglo versus Hispanic samples			Step 2a: Equal loadings				Step 3: Equal covariances		Step 1: EFAs to develop measurement models Step 2b: Partial invariance of factor loadings Step 4: Invariance of structural model Parameters (hypothesized causal effects)
Durvasula, Andrews,	Cross-cultural generalizability	MI as prerequisite to conducting		Step 1: Equivalent	Step 2: Invariant				Step 3: Invariant factor	Step 4: Equal construct means	

Lysonski, and Netemeyer) (1993	of models of attitudes Toward advertising	substantive cross-cultural comparisons		factor structures	factor loadings				covariances		
Li, Harmer, Chi, and Vongjaturapat (1996)	Cross-cultural generalizability of a model of sport behavior	MI as prerequisite to conducting substantive cultural comparisons		Step 2: Pattern invariance	Step 3: Loading invariance	Step 4: Intercept invariance				Step 5: Latent mean invariance	Step 1: EFAs to develop measurement models
Little (1997)	Demonstration of mean and Covariance Structure analysis for Cross-cultural research	MI as prerequisite to conducting substantive cultural comparisons		Step 1: Equivalent factor structures	Step 2: Equivalent factor loadings	Step 3: Equivalent intercepts		Step 4: Equality of latent standard deviations	Step 6: Equality of latent correlations	Step 5: Equality of latent means	Step 7: Invariance of structural model parameters (hypothesized causal effects)
Marín, Gómez, Tschann, and Gregorich (1997)	Condom use among Latino men	MI for men with single versus multiple partners		Step 1: Group specific model	Step 2: Equivalent factor loadings						Step 3 (etc.): Tests of equivalent structural parameters across groups
Palich, Hom, and Griffeth (1995)	Exportability of American management theory and practice	MI as prerequisite to conducting substantive comparisons		Step 1: Constancy of form of the measurement model	Step 2: Equal factor loadings	Step 3: Equal measurement error variances					Step 4: Invariance of structural parameters (hypothesized causal effects)
Riordan and Vandenberg (1994)	Transportability of organizational Measures across cultures	MI as analytic tool to establish equivalence	Step 1: Equivalence of variance/ covariance matrices	Step 2: Test of conceptual equivalence	Step 3: Test of true score equivalence					Step 5: Equivalence of latent means	Step 4: Partially invariant factor loadings
Ryan, Chan, Ployhart, and Slade (1999)	Transportability of surveys in multinational organizations	MI of surveys across countries and languages		Step 1: Equivalent factor patterns	Step 2: Equal factor loadings	Step 3: Equal error variances	Step 4: Equal factor variance-covariances				Step 4: Cross-cultural equivalence in structural model parameters
Singh (1995)	Transportability of management Theories across cultures	Role of MI in transportability of theory		Step 1: Factorial similarity	Step 2: Factorial equivalence						Step 4: Cross-cultural equivalence in structural model parameters

(continued)

Table 2 Continued

Reference	Primary Issue	Specific MI Issue	Measurement Invariance (MI): Invariant Covariance ($\Sigma^g = \Sigma^{g'}$)	Configural Invariance	Metric Invariance ($\Lambda^g = \Lambda^{g'}$)	Scalar Invariance ($\tau^g = \tau^{g'}$)	Invariant Uniquenesses ($\theta^g = \theta^{g'}$)	Structural Invariance (SI): Invariant Factor Variances ($\phi_j^g = \phi_j^{g'}$)	Invariant Factor Covariances ($\phi_{jj'}^g = \phi_{jj'}^{g'}$)	Equal Factor Means ($\kappa^g = \kappa^{g'}$)	Other Tests‘
Cross-cultural comparisons											
te Nijenhuis and van der Flier (1997)	Cross-cultural generalizability of a selection test	MI to test for generalizability	Step 1: Equal covariance matrices	Step 2: Equal factor models	Step 3: Equal factor loadings						
Windle, Iwawaki, and Lerner (1988)	Cross-cultural generalizability of the DOTS-R	MI for American versus Japanese samples			Step 1: Equal factor loading pattern					Step 2: Equality of factor means	
Scale translations											
Byrne and Baron (1994)	Translation of the Beck Depression Inventory (BDI) into French	Equivalence of English and French versions of the BDI		Step 1: Multigroup model with no equality constraints	Step 2: Factor loadings constrained equal						Step 3 (etc.): Partially invariant factor loadings
Smith, Tisak, Bauman, and Green (1991)	Translation of a circadian rhythm Questionnaire into Japanese	Equivalence of the English and Japanese versions			Step 2: Invariant factor pattern						Step 1: EFA to develop measurement models
Cross-group comparisons: Organization/position differences											
Byrne (1991)	Validation of the Maslach Burnout Inventory (MBI) for three levels of educators	MI of the MBI for intermediate,-secondary-, and university-level educators			Step 2a: Noninvariant scaling units				Step 3: Noninvariant factor covariances		Step 1: EFA to develop measurement models separately for groups Step 2b: Partial measurement invariance
Chan (1996)	Cognitive misfit as a component of	MI of cognitive misfit across R&D versus engineering		Step 1: Equivalent factor	Step 2: Factor loadings constrained to						

	Person-environment fit	employee subgroups	patterns	be equal					
Marsh and Byrne (1993)	MTMM analysis of Self-concept data	Within-group (self vs. other rating) and between group (Australians vs. Canadians) MI constraints	Step 1: Baseline MTMM Model					Step 2: Total between group invariance ($\Lambda^g = \Lambda^{g'}$; $\phi_j^g = \phi_j^{g'}$; $\theta^g = \theta^{g'}$ Step 3: Total within-group invariance Step 4: Total within- and between-group invariance	
Marsh and Roche (1996)	Structure of artistic self-concept	MI between performing arts versus other students		Step 2: Factor loadings invariant	Step 5: Uniqueness invariant	Step 3: Factor variance invariant	Step 4: Factor covariances invariant		Step 1: No MI constraints Step 6: $\Lambda^g = \Lambda^{g'}$; $\phi_{jj'}^g = \phi_{jj'}^{g'}$; $\theta^g = \theta^{g'}$ Step 7: Step 6 plus $\phi_j^g = \phi_j^{g'}$
Pike (1996)	MTMM analysis of the viability of using self-Assessments (vs. Objective indicators) of Student achievement	MI of MTMM analysis across 4- versus 2-year higher education institutions	Step 1: Baseline model	Step 2: Invariant factor loadings	Step 4: Invariant uniquenesses		Step 3: Invariant factor covariances		
Van Dyne and LePine (1998)	Multisource performance measurement	MI of self, peer, and supervisor ratings	Step 1a: Equivalent factor pattern loadings across time Step 1b: Equivalent factor pattern loadings across sources	Step 2a: Equivalent factor loadings across time Step 2b: Equivalent factor loadings across sources			Step 3: Invariant factor covariances		

(continued)

Table 2 Continued

Reference	Primary Issue	Specific MI Issue	Measurement Invariance (MI): Invariant Covariance ($\Sigma^g = \Sigma^{g'}$)	Configural Invariance	Metric Invariance ($\Lambda^g = \Lambda^{g'}$)	Scalar Invariance ($\tau^g = \tau^{g'}$)	Invariant Uniquenesses ($\theta^g = \theta^{g'}$)	Structural Invariance (SI): Invariant Factor Variances ($\phi_j^g = \phi_j^{g'}$)	Invariant Factor Covariances ($\phi_{jj'}^g = \phi_{jj'}^{g'}$)	Equal Factor Means ($\kappa^g = \kappa^{g'}$)	Other Tests‘
Cross-group comparisons: Racial/ethnic differences											
Chan (1997)	Race differences of Employment test validity Perceptions	MI to test Black-White differences		Step 1: Equivalent factor patterns	Step 2: Equal factor loadings						
Chan, Schmitt, Sacco, & DeShon (1998)	Examinee reactions to employment testing	MI to test Black-White differences		Step 1: Equivalent factor patterns	Step 2: Equal factor loadings		Step 3: Equal error variances and covariances				Step 4: Equal structural effecte $[\beta'^g : \Gamma'^g] = [\beta'^{g'} : \Gamma'^{g'}]$
Collins and Gleaves (1998)	Genaralizability of the five-Factor Personality model across races	MI to test Black-White differences	Step 1: Equality of covariance matrices	Step 2: Equality of factor form	Step 3: Equality of factor loadings		Step 5: Equality of error matrices	Step 4: Equality of correlations among the factors		Step 6: Equality of latent means	
Tansy and Miller (1997)	Invariance of self-concept	MI of the multidimensional self-concept scale across White and Hispanic populations		Step 1: Equivalent factor patterns	Step 2: Equality of factor loadings		Step 3: Equality of uniquenesses	Step 4: Equality of factor variances and covariances			
Cross-group comparisons: Gender differences											
Byrne (1988)	Generalizability of the SDQ-III across gender	Test of SDQ-III item invariance across gender	Step 1: Null model	Step 2: Baseline model	Step 3: Test of invariance						Step 4 (etc.): Tests of partially invariant factor loadings
Byrne, Baron, and Campbell (1993)	Generalizability of the BDI across gender	Test of BDI item invariance across gender		Step 2: Baseline multigroup model	Step 3: Invariant FOF pattern Step 5: Invariant SOF pattern						Step 1: EFAs to develop measure-ment models Step 4: Test of partially invariant

										FOF loadings
Byrne, Baron, and Campbell (1994)	Gernealizability of the BDI across gender (French Canadians)	Test of BDI item invariance across gender (French Canadians)			Step 2: Equality of factor loadings					Step 1: EFAs to develop measurement models Step 3: Tests of partially invariant factor loadings
Byrne and Shavelson (1987)	Generalizability of the hierarchical self-concept Structure across gender	MI as prerequisite to evaluating substantive hypotheses	Step 2: Invariant variance—covariance matrices	Step 3: Invariant number of factors	Step 4: Invariant loading patterns	Step 7: Series of partially invariant uniqueness	Step 5: Invariant latent variances and covariances			Step 1: EFAs to develop measure ment models Step 6: Partially invariant latent variances and covariances
Li, Harmer, Acock, Vongjaturapat, and Boonverabut (1997)	Validity of the TEOSQ across gender	Test of generalizability of the task and ego orientation in sport questionnaire	Step 1: Equal covariance matrices	Step 2: Equal factor form	Step 3: Equal factor loadings	Step 6: Equal error terms	Step 5: Equal factor variances—covariances		Step 7: Equality of factor means	Step 4: Tests of partially invariant factor loadings
Marsh (1987)	Generalizability of the self-description questionnaire across gender		Step 2: Null model	Step 3: No invariance model	Step 4: Invariant factor loadings	Step 6: Invariant uniquenesses	Step 5: Invariant factor variances and covariances			Step 1: EFA
Stacy, MacKinnon, and Pentz (1993)	Expectancy model of health behavior	MI to test gender differences in measurement and structural properties of the model		Step 1: No cross-group equality constraints	Step 2: Invariant measurement models					Step 3: Invariance of structural model parameters (hypothesized casual effects)
Babcock, Laguna, and Roesch (1997)	MI of processing for different age	speed measures groups		Step 6: Equivalent factor pattern and structure	Step 2 with $\phi_j^g = \phi_j^{g'}$, $\phi_{jj'}^g = \phi_{jj'}^{g'}$, and $\theta^g = \theta^{g'}$ Step 3 with $\phi_j^g = \phi_j^{g'}$ and $\phi_{jj'}^g = \phi_{jj'}^{g'}$ Step 4 with $\theta^g = \theta^{g'}$ Step 5: Invariant factor loadings	Step 2 with $\Lambda^g = \Lambda^{g'}$, $\phi_j^g = \phi_j^{g'}$, and $\phi_{jj'}^g = \phi_{jj'}^{g'}$ Step 4 with $\Lambda^g = \Lambda^{g'}$	Step 2 with $\Lambda^g = \Lambda^{g'}$, $\phi_{jj'}^g = \phi_{jj'}^{g'}$, and $\theta^g = \theta^{g'}$ Step 3 with $\Lambda^g = \Lambda^{g'}$ and $\varphi_{jj'}^g = \varphi_{jj'}^{g'}$	Step 2 with $\Lambda^g = \Lambda^{g'}$, $\phi_j^g = \phi_j^{g'}$, and $\theta^g = \theta^{g'}$ Step 3 with $\Lambda^g = \Lambda^{g'}$ and $\phi_j^g = \phi_j^{g'}$		Step 1: Null model (Σ = I) within each sample

(continued)

Table 2 Continued

Reference	Primary Issue	Specific MI Issue	Measurement Invariance (MI): Invariant Covariance ($\Sigma^g = \Sigma^{g'}$)	Configural Invariance	Metric Invariance ($\Lambda^g = \Lambda^{g'}$)	Scalar Invariance ($\tau^g = \tau^{g'}$)	Invariant Uniquenesses ($\theta^g = \theta^{g'}$)	Structural Invariance (SI): Invariant Factor Variances ($\phi_j^g = \phi_j^{g'}$)	Invariant Factor Covariances ($\phi_{jj'}^g = \phi_{jj'}^{g'}$)	Equal Factor Means ($\kappa^g = \kappa^{g'}$)	Other Tests‘
Cross-group comparisons:											
Gender differences											
Cunningham (1981)	Age-related differences in ability factor structures				Step 2: Invariant factor loadings				Step 3: Invariant factor covariances		Step 1: EFAs to develop measurement models Step 4: Partially invariant factor covariances
Van Ranst and Marcoen (1997)	MI of the Life Regard Index across young and elderly adults			Step 2: Simultaneous baseline model	Step 3: Invariance of factor loadings		Step 5: Invariant error variances	Step 7: Invariance of variances and covariances			Step 1: EFAs to develop measurement models Step 4: Partially invariant factor Loadings Step 6: Partially invariant error variances
Longitudinal/developmental studies											
Hertzog and Schaie (1986)	Stability and change in adult intelligence	MI as a prerequisite to tests of substantive hypotheses		Step 3: Configural invariance only	Step 1: Configural and metric invariance constraints invoked across measurement occasions		Step 7: Residual variances constrained equal across time	Step 4 with $\Lambda^g = \Lambda^{g'}$	Step 5: Factor covariances constrained equal		Step 2: Autocorrelated residuals for identical tests across measurement occasions Step 6: Autocorrelations constrained to be equal with $\Lambda^g = \Lambda^{g'}$, $\phi_j^g \neq \phi_j^{g'}$, and $\phi_{jj'}^g \neq \phi_{jj'}^g$
Pike (1991)	Student growth and development	MI in student outcomes measured longitudinally			Step 3: Lambdas invariant	Step 2: Invariant intercepts in a configurally	Step 4: Epsilons invariant				Step 1: Null baseline model

					Invariant longitudinal measure-ment mode					
Pitts, West, and Tein (1996)	Assessment of stability and change in longitudinal research	MI as a prerequisite to tests of substantive hypotheses	Step 1: Congeneric measurement	Step 2: Tau equivalence		Step 3: Parallel forms				
Schaie, Maitland, Willis, and Intrieri (1998)	Stability of adult psychometric ability factor structures	MI to establish equivalence of measurement models over time	Step 1: Configural invariance	Step 2: Invariance over time Step 3: Invariance across cohorts Step 4: Invariance over time and across cohorts	Step 6: Intercepts invariant over time Step 7: Invariant intercepts across cohorts					Step 5: Partially invariant factor loadings
Multiple applications/dimensions										
Cole, Gondoli, and Peeke (1998)	Examine construct validity of a scale through MTMM	Test structure of the items between sources	Step 1: Equal first-order factors	Step 2: Equal factor loadings			Step 3: Equal factor correlations			Step 4: Equal second-order factor loadings
Hittner (1995)	Examine Construct validity of the TR-AEQ	Examine invariance across gender and across drinking frequency	Step 2a: Congeneric equivalence across gender Step 2b: Congeneric equivalence across drinking frequency	Step 3a: Tau equivalence across gender Step 3b: Tau equivalence across drinking frequency				Step 4a: Factor mean equivalence across gender Step 4b: Factor mean equivalence across drinking f requency		Step 1: Test of single-factor structure using a hold-out sample

(continued)

Table 2 Continued

			Measurement Invariance (MI)					*Structural Invariance (SI)*			
Reference	*Primary Issue*	*Specific MI Issue*	*Invariant Covariance* ($\Sigma^g = \Sigma^{g'}$)	*Configural Invariance*	*Metric Invariance* ($\Lambda^g = \Lambda^{g'}$)	*Scalar Invariance* ($\tau^g = \tau^{g'}$)	*Invariant Uniquenesses* ($\theta^g = \theta^{g'}$)	*Invariant Factor Variances* ($\phi_j^g = \phi_j^{g'}$)	*Invariant Factor Covariances* ($\phi_{jj'}^g = \phi_{jj'}^{g'}$)	*Equal Factor Means* ($\kappa^g = \kappa^{g'}$)	*Other Tests'*
Multiple applications/dimensions											
Hofer, Horn, and Eber (1997)	Examine the robustness and validity of the factor structure underlying the 16PF	Examine measurement properties across gender, police versus felon, and combined		Step 2: Configural invariance conducted on different subgroups	Step 3: Weak tactorial or metric invariance conducted on multiple subgroups	Step 4: Unique means conducted on multiple subgroups; strong factorial invariance	Step 5: Unique variances conducted on multiple subgroups, strict factorial invariance	Step 7: Factor variances conducted on multiple subgroups	Step 6: Factor covariances conducted on multiple sub-groups	Step 7: Factor means conducted on multiple sub-groups	Step 1: Estab-lished factor model in all sam-ples through EFA
Marsh (1993)	Examine the structure of academic self-concept	Examine invariance across gender, four age groups and total sample		Step 2: Factor structure invariance across all eight groups	Step 3: Invariance of factor loadings across all eight groups		Step 6: Invariant uniqueness across all eight groups	Step 5: Invariance of factor variances across all eight groups	Step 4: Invariance of factor correlations across all eight groups		Step 1: Estab-lished factor model in eight groups of gender crossed with four age groups
Nesselroade and Thompson (1995)	Examine selection and other threats to group comparisons	MI as a threat to comparing groups on substantive measures	Step 1: Baseline model conducted twice; twins and ability groups	Step 6: Same factor pattern	Step 5: Equal factor loadings		Step 4: Equal unique variances; also with $\Lambda^g = \Lambda^{g'}$		Step 3: Equal factor covariances also with $\Lambda^g = \Lambda^{g'}$, $\tau^g = \tau^{g'}$, $\theta^g = \theta^{g'}$, and $\phi_{jj'}^g = \phi_{jj'}^{g'}$	Step 2: Equal common factor means also with $\Lambda^g = \Lambda^{g'}$, $\tau^g = \tau^{g'}$, $\theta^g = \theta^{g'}$, $\phi_j^g = \phi_j^{g'}$, and $\phi_{jj'}^g = \phi_{jj'}^{g'}$	All steps conducted twice;

Parker, Baltes, and Christiansen (1997)	Examine moderating effects of gender and ethnicity on Perceptions of Organizations' support for AA/EEO	MI as a prerequisite to substantive comparisons across gender and race	Step 1: Same pattern	Step 2: Weak invariance	Step 3: Strong invariance				Step 4: Control variables invariant
Schulenberg, Shimizu, Vondracek, and Hostetler (1988)	Generalizability of CDS across grade level and across gender	Used as primary tool for examining generalizability	Step 2: Equal factor patterns	Step 3: Equal factor loading or metric invariance		Step 5: Equal unique variances	Step 4: Equal factor variances and covariances		Step 1: Series of EFAs to establish factor model
Van den Bergh and Van Ranst (1998)	Generalizability of self-perception profile across gender and across age	MI as primary tool for examining generalizability	Step 2: Simultaneous baseline model	Step 3: Invariance of factor loadings		Step 4: Invariance of error variances		Step 5: Invariance of factor correlations	Step 1: Established separate baseline models in each group
Veerman, ten Brink, Straathof, and Treffers (1996)	Generalizability of SPPC across nonclinical and clinical Groups	MI as primary tool for examining generalizability	Step 2: Simultaneous baseline	Step 3b: Equivalent factor loadings		Step 4: Equivalent accuracy		Step 3a: Equivalent structure	Step 1: Established separate baseline models in each subgroup

ME/I has been investigated in longitudinal studies in conjunction with human developmental processes (e.g., Pitts, West, & Tein, 1996) or other changes (e.g., socialization) (see Chan, 1998).

Generally speaking, Table 2 shows that the frequency with which a test was undertaken varied considerably from test to test. For example, although the majority of the studies reported tests of configural invariance (59/67 or 88%) and metric invariance (i.e., $\Lambda^g = \Lambda^{g'}$, 66/67 or 99%), they less often reported tests of invariant unique variances (i.e., $\Theta^g = \Theta^{g'}$, 33/67 or 49%), tests of invariant factor variances and covariances ($\Phi_j^g = \Phi_j^{g'}$, 22/67 or 33%, and $\Phi_{jj'}^g = \Phi_{jj'}^{g'}$, 39/67 or 58%, respectively), or omnibus tests of invariant observed variables' covariance matrices (i.e., $\Sigma^g = \Sigma^{g'}$, 15/67 or 26%). Still fewer studies conducted tests of latent mean differences (i.e., $\kappa^g = \kappa^{g'}$, 14/67 or 21%) or tests of scalar invariance (i.e., $\tau^g = \tau^{g'}$, 8/67 or 12%). Thus, as reflected in Table 1 as well, the overwhelming concern among sources in Table 2 was establishing configural and metric invariance (i.e., conceptual equivalence of measures and equivalent calibration of measures to constructs across groups, respectively). It is difficult to state why researchers may not have proceeded beyond the test for metric invariance. For one, they could have stopped because results indicated that they should not proceed further. Second, researchers may have been unaware that other tests could have been undertaken. Third, even if they had known about some of the other tests, the software for conducting them may not have existed or been as flexible as it is today to undertake these tests. Finally, some would not have been able to do so because they did not incorporate the vector of means in the analysis. Without this vector, one cannot conduct a test for scalar invariance.

Second, we found it odd that although 62% of the papers listed in Table 1 recommended starting ME/I tests with the omnibus test of differences in covariance matrices ($\Sigma^g = \Sigma^{g'}$), only 22% of the applications in Table 2 actually did so. Thus, most studies listed in Table 2 undertook tests of specific aspects of ME/I without first determining whether measures were invariant overall. Third, there were inconsistencies in the manner and sequencing of tests for invariant uniqueness (i.e., $\Theta^g = \Theta^{g'}$). This is important because (a) nearly half of the studies (33 of the 67) undertook this test, and (b) it is often undertaken as a test of invariant reliabilities across groups. The equality restriction that $\Theta^g = \Theta^{g'}$ (as compared to a less restricted model that does not invoke these constraints) is an appropriate test of invariant unique variances (and possibly covariances) across groups. It may be appropriately invoked as a test for invariant reliabilities only if factor variances are invariant across groups (i.e., $\phi_j^g = \phi_j^{g'}$) (see Cole & Maxwell, 1985; Rock et al., 1978). Otherwise, invariance in measures' reliabilities must be tested by adjusting for group differences in factor variances:

$$\phi_j^g / (\phi_j^g + \theta_{jk}^g) = \phi_j^{g'} / (\phi_j^{g'} + \theta_{jk}^{g'}), \tag{5}$$

where $\phi_j^g \neq \phi_j^{g'}$ (i.e., no invariance constraints are effected on factor variances across groups), and the hypothesis of equal reliabilities is tested by comparing alternate models in which uniquenesses are constrained to be equal ($\Theta_{jk}^g = \Theta_{jk}^{g'}$) and in which they are not ($\Theta_{jk}^g = !\Theta_{jk}^{g'}$). Thus, if tests of invariant uniqueness are conducted with the intent of testing equal reliabilities across groups, a test of the invariance of factor vari-

ances must be effected first. However, as Table 2 shows, only 52% (17/33) of the studies that tested for invariant uniqueness also tested for invariant factor variances, and all of these tested for invariant factor variances after conducting tests of invariant uniqueness.

Fourth, there was little consistency in the number or combination of ME/I tests conducted in the studies listed in Table 2. This is understandable to the extent that the ME/I tests that were effected addressed theoretical concerns that themselves varied from study to study. No study undertook all eight ME/I tests (or even seven of the eight tests) listed at the top of Table 2. Most often, studies reported two to four of these eight ME/I tests. Of these, tests for configural and metric invariance were most often reported, with little consistency across studies in terms of which additional tests were conducted.

Fifth, as reflected in Table 1 as well, there was some consistency in terms of the sequencing of steps in evaluating ME/I. Specifically, tests of configural invariance were generally conducted before all other tests in a sequence (although the sequence that followed varied considerably). Also, the test of configural invariance was often followed by a test for metric invariance. However, beyond this, there was little consistency across studies with respect to the sequencing of steps. This may have been due to different substantive issues, which called for the use of different invariance tests. Two issues that are reflected in the final column of Table 2 ("Other Tests") complicated our ability to discern exactly how tests were sequenced. First, some studies (e.g., Babyak, Snyder, & Yoshinobu, 1993; Byrne, 1994; Pentz & Chou, 1994) reported completing exploratory factor analyses (EFAs) on measures prior to conducting CFA-based ME/I tests. Although this practice may be justified in the earlier stages of scale development (e.g., Pentz & Chou, 1994), we feel it is inappropriate to conduct EFAs and CFA-based tests of ME/I using the same sample data and for "established" measures. The second issue concerned tests of partial invariance. Partial invariance was tested primarily in conjunction with tests for metric invariance (e.g., Byrne, 1991; Finch, Okun, Barrera, Zautra, & Reich, 1989; Li, Harmer, Acock, Vongjaturapat, & Boonverabut, 1997; Riordan & Vandenberg, 1994) but was invoked for other ME/I tests as well (e.g., Dumka, Stoerzinger, Jackson, & Roosa, 1996, applied it to structural parameters; Hattrup, Schmitt, & Landis, 1992, applied it to tests of invariant uniquenesses). As mentioned in our review of Table 1, tests of partial invariance are conducted with the ideas that (a) measures may be invariant across some but not all groups or that some but not all of the measure's components (e.g., items) are invariant across groups, and (b) implementing controls for partial measurement inequivalence renders permissible cross-group comparisons that might otherwise not be appropriate. Note that tests of partial invariance with respect to structural or measurement parameters other than factor loadings are meaningful only to the extent that configural invariance and (at least partial) metric invariance have been established previously. In practice, however, tests for partial ME/I are often conducted in an exploratory fashion and iteratively, so that the entire sequence of ME/I tests is not conducted on a strict stepwise basis.

Finally, there was little consistency in terms used to describe the substantive hypothesis tested by each of the ME/I tests conducted, and these differences in nomenclature were even greater for the empirical studies in Table 2 than they were for the proposed practices in Table 1. For example, what we have chosen to refer to as a test of

configural invariance was referred to variously as "baseline model," "invariant number of factors," "test of form," "pattern invariance," "factorial similarity," and so on. Similarly, what we referred to as invariant uniqueness was referred to by authors as "equivalence of error structure," "equivalence of first-order factor unique variance terms," "strict factorial invariance," "equality of error matrices," and so on. The nomenclature varied considerably across studies for all ME/I tests and usually reflected the authors' particular substantive concerns (e.g., alpha, beta, and gamma change) (Schmitt, 1982; Vandenberg & Self, 1993). Nomenclature was seldom inappropriate, but it was equally seldom common across studies (e.g., one author's "baseline model" was another author's "test of form"). This reinforces even further the need to adopt the common nomenclature that we present at the top of Tables 1 and 2.

A Closer Look at Each Test of ME/I

Up to this point, we have (a) argued that tests of ME/I are important logical prerequisites to conducting cross-group comparisons, (b) shown that a number of specific aspects to ME/I are testable within a CFA framework, (c) shown that ME/I issues have been of concern to researchers in a variety of disciplines, and (d) suggested that there is little consistency in recommended tests for ME/I or in their application in empirical research. In this section, we take a closer look at each of the eight ME/I tests listed in Tables 1 and 2. There are three reasons for doing so. First, we wish to state explicitly the logic and interpretation of each step. Second, our review of the articles in Tables 1 and 2 uncovered important details associated with each test that should be considered in their applications. Finally, the following section acts as a backdrop for our proposed recommendations for effecting tests of ME/I. Our discussion is organized according to the presentation of each test along the tops of Tables 1 and 2.

$\Sigma^g = \Sigma^{g'}$. This test is typically effected in a multisample application of CFA by testing for equality of samples' covariance matrices, that is, $\mathbf{S}^g$ versus $\mathbf{S}^{g'}$. The tenability of the null hypothesis is evaluated through the chi-square statistic and other overall goodness-of-fit indices. Failure to reject the null hypothesis that $\Sigma^g = \Sigma^{g'}$ is commonly viewed as a demonstration of overall measurement equivalence across groups. Consequently, further tests of specific aspects of ME/I are neither needed nor warranted. On the other hand, rejection of the null hypothesis is a preliminary indication that some form of ME/I exists between groups (Schaubroeck & Green, 1989; Schmitt, 1982; Vandenberg & Self, 1993). However, rejection of the null hypothesis that $\Sigma^g = \Sigma^{g'}$ is uninformative with respect to the particular source of measurement inequivalence and thus "argues for the testing of a series of increasingly restrictive hypotheses in order to identify the source of nonequivalence" (Byrne, 1989, p. 126). The "series of increasingly restrictive hypotheses" may be in the form of one or more of the other ME/I tests.

Test of configural invariance. As noted earlier, this is a test of the null hypothesis that the a priori pattern of free and fixed factor loadings imposed on the measures' components (e.g., items) is equivalent across groups (Horn & McArdle, 1992). This test may also be thought of in terms of evaluating the hypothesis of invariant congeneric measurement properties across groups. Inasmuch as the factor structure is a reasonable empirical map of the underlying conceptual or cognitive frame of reference used to make item responses (Vandenberg & Self, 1993), then the differences between

groups (if they exist) are evidenced by different factor structures across groups (e.g., three factors characterize responses of one group, whereas different numbers of factors or a different factor pattern configuration characterize some other group or groups). Failure to reject the null hypothesis (i.e., finding support for ME/I) has two implications. First, it means that respondent groups were employing the same conceptual frame of reference and thus ultimately might be compared (e.g., in tests of latent mean group differences) with reference to measures that reflect equivalent underlying constructs. Second, it means that the further tests of additional aspects of ME/I may proceed inasmuch as they are nested within the test of configural invariance. However, if the null hypothesis is rejected, neither tests of group differences (e.g., tests of latent mean group differences, group differences in structural parameters) nor additional ME/I tests are justified—it makes no sense to conduct tests of group differences when the constructs that are being measured differ across groups. Tests for additional aspects of ME/I are similarly equally problematic. It is meaningless, for example, to test whether like items are calibrated to the constructs (ξ_k) equivalently between groups when the underlying constructs themselves are different from group to group.

$\Lambda_X^g = \Lambda_X^{g'}$. This is the test of metric invariance between groups (Horn & McArdle, 1992) and is effected by constraining the factor loading (λ_{ik}) of like items to be equal across groups. This is a stronger test of factorial invariance than is the test of configural invariance in that in addition to specifying an invariant factor pattern, loadings of like items within that pattern are now constrained to be equal. Factor loadings are the regression slopes relating the X_{jk} to their corresponding latent variables, ξ_j (Bollen, 1989), and thus represent the expected change in the observed score on the item per unit change on the latent variable. Thus, the test of the null hypothesis that $\Lambda_X^g = \Lambda_X^{g'}$ is a test of equality of scaling units across groups (Jöreskog, 1969; Schmitt, 1982; Vandenberg & Self, 1993).

Studies in Tables 1 and 2 were in disagreement, however, as to the ramifications of rejecting the null hypothesis. Some authors (e.g., Bollen, 1989; Millsap & Hartog, 1988) believe that the rejection of the null hypothesis should preclude further tests of ME/I much like the rejection of the null hypothesis of configural invariance should preclude a test of metric invariance. Others, however (e.g., Byrne et al., 1989; Steenkamp & Baumgartner, 1998), have argued for investigation of partial metric invariance in the event that the overall test of the null hypothesis of metric invariance is rejected. Inasmuch as they were presented earlier in the discussions of the tables, we will not repeat the rationale underlying partial metric invariance here. We would, however, like to make more explicit the relative advantages and disadvantages of testing for partial metric invariance given its frequency of use. First, statistical criteria (or other goodness-of-fit heuristics) for relaxing metric invariance constraints have not been applied consistently in the literature (see Byrne, 1991; Byrne & Baron, 1994; Byrne, Baron, & Campbell, 1993, 1994; Dumka et al., 1996; Finch et al., 1998; Li et al., 1997; Marsh & Hocevar, 1985; Pentz & Chou, 1994), and there continues to be no consensus in their application. Second, invoking partial invariance constraints, as applied to date, generally has been an exploratory, iterative, post hoc practice, and so it is subject to capitalization on chance. On the other hand, specification of partial invariance constraint effects control for measurement inequivalence for those indicators that do not satisfy metric invariance constraints across groups and thus allows further tests of additional aspects of ME/I.

Although we recognize the value in controlling for partial invariance to facilitate the evaluation of more substantive research hypotheses, we also recognize the threats that exploratory, data-driven analyses pose to the integrity of research findings. Thus, we recommend a conservative approach to implementing partial invariance constraints. Although Steenkamp and Baumgartner (1998) indicate that metric invariance constraints can be relaxed up to the point that only the reference indicator and one other indicator retain them (see p. 81), we recommend that metric invariance constraints be relaxed (a) only for a minority of indicators, (b) on as strong a theoretical basis as is possible, and (c) when cross-validation evidence points to their viability. Alternately, indicators that do not meet metric invariance restrictions may be removed from analysis.

$\tau_X^g = \tau_X^{g'}$. As noted earlier, this test of the hypothesis that the vector of item intercepts is invariant (scalar invariance) across groups was one of the least frequently conducted tests. Inasmuch as the τ vector contains location parameters for the X_{jk} (i.e., the value of an observed item when the value of the corresponding ξ_j is zero), this test has been interpreted by some as a test for systematic response bias (e.g., leniency) differences between the groups (Bollen, 1989) for comparisons in which latent mean group differences are not otherwise expected. For example, differences in rating items' intercepts across rater source groups would indicate a leniency bias, assuming that each rater source is privy to the same sets of ratee performance-related behaviors. On the other hand, intercept differences may not reflect biases (undesirable) but response threshold differences that might be predicted based on known group differences (desirable), for example, between inexperienced versus highly experienced employees. Thus, whether this invariance test should be undertaken depends greatly on the substantive context underlying the study. For example, assume that after conducting the tests for invariance, substantive reasons will result in one group (males) being compared with another group (females) where it is hypothesized that one group should have a higher mean on the construct of interest than the other group. Assuming also that the measure is a valid operationalization of the construct and that the hypothesis regarding group differences is true, then the items underlying that measure should also reflect group differences if mean difference tests were conducted on an item-by-item basis. Hence, a test for intercept or scalar invariance (i.e., no differences between groups) is not appropriate because difference in item location parameters would be fully expected. However, these differences are not biases in the sense of being undesirable as in rating source biases, but rather they reflect expected group differences.

$\Theta_{\delta_g} = \Theta_\delta^{g'}$. As noted previously, this is a test of the invariance of the unique variances (and possibly covariances) across groups. Ordinarily, Θ_δ is specified as diagonal so that indicators' uniquenesses are assumed uncorrelated. However, covariances among uniquenesses are sometimes estimated, for example, to control for item-specific effects as they may covary over repeated measures (e.g., Chan, 1998; Hertzog & Schaie, 1986; Vandenberg & Self, 1993). In either case, this test is undertaken by constraining like items' uniquenesses (δ_{jk}) to be equal between groups. Likewise, uniqueness covariances may be constrained to be equal between adjacent measurement occasions to test the plausibility of a first-order autocorrelated structure on the uniquenesses (e.g., Willett & Sayer, 1994).

The test that $\Theta_{\delta_g} = \Theta_{\delta}^{g'}$ has been treated by most researchers as a test for invariant indicator reliabilities across groups (e.g., Schmitt et al., 1984). However, as we noted earlier, this is a proper test of invariant reliabilities only if it has been established first that $\Theta_{j_g} = \Theta_{j}^{g'}$ (i.e., that the factor variances are invariant across groups). Otherwise, (a) the test that $\Theta_{\delta_g} = \Theta_{\delta}^{g'}$ should be properly interpreted as a test of invariant unique variances, or (b) the constraint described in Equation (5) should be implemented to test for invariant reliabilities.

$\Phi_{j}^{g} = \Phi_{j}^{g'}$. This is a test of the hypothesis that factor (i.e., ξ_j) variances are invariant across groups and is invoked by placing equality constraints between like factor variances (i.e., the diagonal elements of Φ) across groups. Factor variances represent the dispersion of the latent variables (ξ_j) and thus represent variability of the construct continua within groups. Failure to reject the null hypothesis that $\Phi_{j}^{g} = \Phi_{j}^{g'}$ indicates that groups used equivalent ranges of the construct continuum to respond to the indicators reflecting the construct(s). Rejection of the null hypothesis indicates that the group with the smaller factor variance is using a narrower range of the construct continuum than is the group with the larger factor variance. Our review indicated that this test has been used primarily in two contexts. The first (as discussed in the previous paragraph) was as a precondition to testing the invariance of item reliabilities. The second (in conjunction with a test of metric invariance) was as an additional test for beta change (e.g., Schaubroeck & Green, 1989; Schmitt, 1982; Vandenberg & Self, 1993). In the latter contexts, these tests were interpreted as providing complementary factor-level and item-level assessments of differential scale calibration, respectively.

$\Phi_{jj}^{g} = \Phi_{jj}^{g'}$. This is a test of the hypothesis of equal factor covariances across groups and is accomplished by constraining the covariances of like factor pairs to be equal across groups (e.g., $\varphi_{21}^{g} = \varphi_{21}^{g'}$, etc.). This test has most often been applied to test for equality of factor intercorrelations (by combining this and the previous test into one simultaneous test) (see Byrne & Shavelson, 1987; Cole, Gondoli, & Peeke, 1998; Collins & Gleaves, 1998; Finch & West, 1997; Li et al., 1997; Marsh, 1987; Pentz & Chou, 1994; Tansey & Miller, 1997; Van Ranst & Marcoen, 1997). Less frequently, it was used as a second test of gamma change (in conjunction with a test of configural invariance) (see Schaubroeck & Green, 1989; Schmitt, 1982; Schmitt et al., 1984; Vandenberg & Self, 1993). Of the two applications, researchers in the first group were unclear in most cases explaining the rationale and implications of undertaking the test. Although less frequent, researchers in the second group were very clear as to why they were undertaking the test. Specifically, if the conceptual domain is constant across groups (i.e., responses to items are driven by the same conceptual frame of reference), then the relationships among the factors should also not differ between groups. If, on the other hand, the conceptual domain is not constant, then the covariances among the factors may differ (and perhaps dramatically) in terms of both strength and pattern.

Nevertheless, we think there is little to be gained by applying this test to evaluate constancy of the conceptual domain for the following reasons. First, if the null hypothesis of configural invariance has been rejected previously, then the number, strength, and pattern of factor covariances likely also vary across groups. Consequently, the test is redundant. Second, if the null hypothesis of configural invariance has not been rejected previously but the null hypothesis of invariant factor covariances

is, we would be hard-pressed to claim that the conceptual domain has changed on the basis of the latter test alone when the more stringent test (i.e., configural invariance) indicated otherwise.

$\kappa^{g} = \kappa^{g'}$. This is a test of equal factor means across groups. Rather than testing some property of the underlying scale, it is the desired substantive test to determine whether (experimentally assigned or intact) groups differ in level on the underlying construct(s) ξ_j that are operationalized (and approximated) by the composite of the X_{jk}s. Like the more traditional analyses (ANOVA, *t* test), the test of latent mean differences begins with an omnibus test of the null hypothesis by constraining the means to be equal across all groups. If the null hypothesis is rejected, then subsequent tests may be conducted to isolate specific differences between groups (e.g., see Riordan & Vandenberg, 1994; Schaubroeck & Green, 1989; Vandenberg & Self, 1993). Nevertheless, just as in more traditional post hoc comparisons, alpha levels should be adjusted for multiple comparisons among latent means to preserve the desired overall Type I error rate.

So, why not stick with these more traditional analytic approaches? There are four reasons. First, they may not be appropriate. As we argued earlier in this article, if substantial measurement inequivalence exists across groups, it is inappropriate to compare mean group differences on nonequivalent measures. Comparisons of apples to apples are meaningful. Comparisons of apples to sandwiches to sand wedges are not. Second, even if measurement equivalence incurs, tests that are more traditional are rarely (if ever) preceded by tests of ME/I. Thus, tests of ME/I that support measurement equivalence provide the justification for testing for group differences, whether these tests are conducted on observed measures or between latent means. Third, tests of differences between latent means are disattenuated for measurement error. Corrections for attenuation due to unreliability are rarely (if ever) effected in more traditional tests (Schmidt & Hunter, 1996). Finally, control for partial measurement inequivalence across groups can be effected by implementing partial invariance constraints (Byrne et al., 1989). This is just not possible in more traditional analytic approaches (Cole, Maxwell, Arvey, & Salas, 1993).

Special issues. Several additional, specific issues regarding ME/I deserve mention that did not fit neatly within our more general review of the ME/I literature. We mention a few here. One concerns the form of the data input for analysis. Tests of ME/I issues should be conducted within an analysis of covariance and mean structures framework, and much of our discussion was with respect to multisample analyses within this framework. Most of the research we reviewed did and must test for ME/I using a multisample approach (e.g., tests of independent group differences, cross-cultural comparisons, between-groups experimental manipulations). However, some longitudinal studies included in our review input a single augmented covariance matrix for analysis (e.g., Hertzog & Schaie, 1986; Vandenberg & Self, 1993). In these studies, the augmented matrix contained covariances between measures across all time frames over which ME/I issues were of interest. That is, the augmented covariance matrix takes into account the lagged associations between measured variables as well as the within-time covariances. Thus, ME/I issues can be tested in time-structured

(i.e., multiwave longitudinal) data using either the multisample approach or an augmented covariance matrix as input (we illustrate the latter option in our example later). There are advantages and disadvantages to both. The primary advantages of using the augmented covariance matrix are that (a) other aspects of change, in addition to ME/I issues, can also be assessed (we will return to this shortly), and (b) controls for autocorrelations among measures' specificities (e.g., a first-order autocorrelated "error" structure) can be effected that cannot be implemented using the multisample approach. The major disadvantages, both of which stem from input of a (typically much) larger data array, are (a) increased likelihood of nonconvergent or improper solutions and (b) generally worse model fit due to the availability of greater model degrees of freedom. Generally speaking, we recommend analysis of the augmented covariance matrix when possible as it takes advantage of the complete data structure.

Related to the above, a number of studies investigated ME/I issues as one aspect of longitudinal change (e.g., Bartunek & Franzak, 1988; Hertzog & Schaie, 1986; Pike, 1991; Schaie, Maitland, Willis, & Intrieri, 1998; Schaubroeck & Green, 1989; Vandenberg & Self, 1993). The measurement of change has been a longstanding and controversial topic (Cronbach & Furby, 1971; Lance et al., in press). One approach to the assessment of change in a structural equation modeling framework is latent growth modeling (LGM) (Duncan & Duncan, 1995; Duncan et al., 1994; McArdle, 1988; McArdle & Aber, 1990). One of the basic aims of LGM is to identify latent initial status and change variables (LGM allows for various functional forms of change) from observed measures tracked longitudinally over at least three measurement waves. Interest usually focuses on the mean growth trajectory across subjects, individual differences in growth trajectories, and predictability of individuals' growth (or decline) parameters based on individual differences. Based on the present discussion, it would seem that establishment of longitudinal ME/I would be a natural prerequisite to assessment of change using LGM, yet curiously, the issue has received scant attention in this context (however, see Chan, 1998; Lance, Vandenberg, & Self, 1999). We urge researchers to consider ME/I issues prior to conducting longitudinal change assessment. ME/I is just as important a prerequisite for the assessment of change over time as it is for conducting comparisons between (nonequivalent) groups' means.

Third, much of our presentation has been based on the idea that manifest indicators of latent variables are at their lowest level of aggregation (i.e., individual scale items). As such, our presentation has been at the level of what Bagozzi and Heatherton (1994) refer to as a "total disaggregation" model (see also Bagozzi & Edwards, 1998), and this is the usual practice in ME/I research. However, researchers can and do form manifest indicators at higher levels of aggregation, for example, by forming subscales or "testlets" (a "partial aggregation" model) or a single composite indicator (a "total aggregation" model) (Bagozzi & Heatherton, 1994). One interesting extension to the ME/I literature is the investigation of ME/I issues at the subscale level of aggregation (Labouvie & Ruetsch, 1995). Although not uncontroversial (see the series of comments on Labouvie & Ruetsch, 1995, and the authors' rejoinder), this represents an extension to the ME/I literature that addresses the manner in which many researchers actually do construct manifest indicators for theoretical variables of substantive interest.

Fourth, we have discussed ME/I issues only as they relate to connections between manifest indicators X_{jk} and the latent constructs ξ_j that they are intended to reflect. This also is consistent with the ME/I literature. However, with the exception of modeling autocorrelated specificities, we have not discussed effects of measurement bias. As Campbell and Fiske (1959) pointed out more than 40 years ago, manifest variables likely reflect some aspects of the methods employed in the measurement process as well as the intended underlying construct. Modeling autocorrelated error structures in analyses of longitudinal data using augmented covariance matrices partly addresses this issue. Marsh and Byrne (1993) suggest another approach that combines features of the analysis of multitrait-multimethod data and tests of ME/I issues. They show how hypotheses concerning the invariance of (a) both trait and method factor loadings across populations and (b) trait factor loadings across measurement sources are testable. As such, their proposal extends tests of ME/I to issues of measurement method bias as well as more traditional issues of construct validity.

Fifth, tests for scalar invariance and latent mean differences cannot be viewed as being independent because they are, in fact, closely related (G. Cheung, personal communication, January 11, 2000). The relationship incurs due to identification requirements for the latent means. For example, assume that one is interested in comparing latent means across two groups on a K-item unidimensional scale. In this scenario, there are $2 \bullet K$ observed means (i.e., the K-items means in each of the two groups) but $2 \bullet K + 2$ unknowns (the K-items' intercepts in each of the two groups plus each of the group's latent means). Consequently, the model is underidentified in terms of its location parameters. Identification is usually achieved in one of two ways. In the gth group, the expected value of the kth item can be expressed as $E(X_k) = \tau_k + \lambda_k\xi$ (because $E[\delta_k] = 0$) (Cole et al., 1993) and suggests that the latent mean (and, consequently, the model as a whole) can be identified by fixing one item's (e.g., the reference item's) intercept equal to zero in each group (Bollen, 1989; Steenkamp & Baumgartner, 1998). Thus, the reference item in the gth group is defined as $E(X_k) = 0.0 + \lambda_k\xi$. But this solution to the identification problem is predicated on (a) prior establishment of metric invariance (i.e., the $\lambda_k^g = \lambda_k^{g'}$) as indicated earlier and (b) the (untested) assumption of scalar invariance for the reference item (i.e., $\tau^g = \tau^{g'}$ for the reference item). If the latter assumption holds, then the latent means are inextricably (and appropriately) tied to a function (i.e., weighted by λ_k^g of the referent items' means). If not, tests of latent mean differences are confounded with noninvariant referent items' intercepts because items' means are functions of both the latent means and the items' regression intercepts. The issues are similar in another common approach to identification—namely, fixing the mean of one latent variable (κ^g) to zero and imposing equality constraints for the reference items' intercepts across groups (Jöreskog & Sörbom, 1996; Steenkamp & Baumgartner, 1998). The key issue here that, to our knowledge, has not been suitably resolved for tests of ME/I is proper selection of scalar-invariant reference indicators to identify latent means and as a prerequisite to conducting tests of latent mean differences.

Finally, there is the issue of assessing model fit. Obviously, a researcher's judgments concerning the appropriateness of imposed invariance constraints depend on the fit indices associated with a specific model and/or the differences between hierarchically nested models. Not so obvious, though, is that the general issue of fit (i.e., which index to use, what level of an index indicates acceptable model fit, etc.) is, simply

stated, in a state of evolution. The past several years have witnessed a number of published (e.g., Cheung & Rensvold, 1999b; Hu & Bentler, 1998, 1999; Marsh, Balla, & Hau, 1996; Tanaka, 1993) and unpublished manuscripts (e.g., Cheung & Rensvold, 1999a, 1999c) on the topic of model fit, and like all evolving research areas, both good news and bad news emerge from this research. The bad news is that despite our desire to possess a set of critical values against which we can make a definitive "fit" or "no-fit" decision, none is unambiguously forthcoming from the literature. Indeed, current research indicates that even some of the most widely accepted values for certain indices (e.g., .90 or above for the Tucker-Lewis index) may not have been appropriate after all (Hu & Bentler, 1999, show that .95 may be more appropriate). The good news, though, is that these researchers are converging on some general agreements and are addressing critical areas for improving our use of fit indices to make valid decisions regarding model fit.

In the following paragraphs, we summarize recent literature on model goodness of fit as it relates to judging the appropriateness of invariance constraints. As seen shortly, this is still a complex topic. Our brief review is organized into three sections: (a) overall model fit, (b) differences between models, and (c) reference indicator selection and partial invariance models.

Overall model fit. Overall model fit refers to evaluating the ability of the a priori model to (at least approximately) reproduce the observed covariance matrix. Overall model fit has its largest role in appraising the test for configural invariance. Recall that a good overall fit results in deciding that the factor structure (i.e., configuration) underlying a set of measures is equivalent from one group to the next. In contrast, a poor fit means that the measures were not anchored to the same configuration of latent variables in each of the comparison groups. In the former "good-fit" case, it is permissible to continue with further invariance tests, but in the latter "poor-fit" case, further testing is not appropriate.

Among the researchable fit issues, researchers have concentrated most heavily on the topic of overall fit (e.g., Bollen, 1989; Hu & Bentler, 1998, 1999; Marsh et al., 1996; Marsh, Balla, & McDonald, 1988; Medsker, Williams, & Holahan, 1994; Mulaik et al., 1989; Tanaka, 1993). Two characteristics common to most of the studies are (a) evaluating the impact of variations in model, study, and sample characteristics (e.g., sample size, model misspecification, etc.) on the values of the fit indices and (b) using the evaluation to make recommendations as to which fit indices are most appropriate (e.g., Tanaka, 1993, p. 32, Table 2.2). Three recommendations have appeared from this research literature. The first is to interpret the chi-square test of model fit only in conjunction with other practical fit indices (Bollen, 1989; Bollen & Long, 1993; Marsh et al., 1996; Medsker et al., 1994). Although ideally, a statistically nonsignificant chi-square value should be observed to infer support for a well-fitting model, it is a fact that a statistically significant chi-square value can incur even though there are only minor differences between the groups' factor patterns. This results from the sensitivity of the chi-square test to even minor deviations between the groups' sample covariance matrices and the chi-square test's susceptibility to sample size influences (Bollen & Long, 1993; Hu & Bentler, 1993; James et al., 1982).

The second recommendation is to select a variety of practical fit indices with which to supplement the chi-square test and not just one. The advice emerging from the

research literature, though, varies with respect to which of the practical fit indices to use and what critical value of the index needs to be met to infer good model fit. The variation is due to the trade-offs in the adoptions of each fit index, with no single index emerging as one to embody all trade-offs. Although we recommend four such indices, we do with the additional recommendation that researchers only adopt them after having themselves reviewed the relevant literature. Our four recommendations are the following: (a) Tucker-Lewis index (TLI) (Tucker & Lewis, 1973), now often referred to as the nonnormed fit index (NNFI); (b) relative noncentrality index (RNI) (McDonald & Marsh, 1990); (c) root mean square error of approximation (RMSEA) (Steiger, 1990); and (d) the standardized root mean square residual (SRMR) (Bentler, 1995). Selection of the TLI and RNI was based on the work by Marsh et al. (1996) and Hu and Bentler (1998, 1999). Unlike other fit indices, the TLI and RNI were not systematically related to sample size, and both reflected systematic variation in model misspecifications, particularly misspecified factor loadings. The difference between the TLI and RNI was that the TLI but not the RNI also appropriately penalized model complexity (thereby appropriately rewarding model parsimony). This property of the TLI is particularly useful in tests of nested models in which the imposition of equality constraints is used to test the invariance of solutions across multiple groups (Marsh, 1995; Marsh et al., 1996). Until recently, researchers agreed that TLI and RNI values of .90 or above were indicative of well-fitting models. Hu and Bentler (1999) demonstrated recently, though, that the prior advice was based on studies that did not adequately account for variations in model misspecifications. Based on their simulation, Hu and Bentler (1999) present evidence that only when observed TLI and RNI values equal or exceed .95 should an inference of "good fit" be used to describe a model. However, Hu and Bentler's (1999) study is one of the first of its kind and is in need of extension investigating additional simulation design characteristics. Until these extensions are completed and their findings supported, it may be premature to throw out the .90 critical value. Their study, however, indicates that the .90 should not be viewed as the de facto criterion. Our recommendation at this juncture is to view the .90 as a lower bound of good fit, with high confidence in fit emerging when the TLI and RNI meet or exceed .95. Again, this is an area that warrants consistent monitoring, and no researcher should adopt these standards without first examining the current literature.

Unlike the TLI and RNI, the RMSEA (our third recommendation) does not require a null model in its calculation and does not conflict with the requirements for parsimony (Browne & Cudeck, 1993; Jöreskog & Sörbom, 1996). Like the TLI and RNI, though, the RMSEA is also sensitive to model misspecifications and especially misspecified factor loadings (Hu & Bentler, 1998, 1999). The RMSEA addresses, "How well would the model, with unknown, but optimally chosen parameter values fit the population covariance matrix if it were available?" (Browne & Cudeck, 1993, pp. 237-238). Ideally, there should be no error, but realistically, "values up to .08 represent reasonable errors of approximation in the population" (Jöreskog & Sörbom, 1996, p. 124). For the same reasons stated above, Hu and Bentler (1999) recently challenged the latter value and, based on their findings, stated that a critical value of .06 or less was most likely to prevent the acceptance of truly misspecified models. We would add, though, that again this recommendation is coming from a single study and that until more studies are completed, the value of .08 is not unreasonable but should perhaps be looked on as an upper limit.

The final practical fit index, the SRMR, was selected because of recent evidence presented by Hu and Bentler (1998, 1999). Namely, unlike the latter three indices, the SRMR was most sensitive to model misspecifications among the factor covariances. As with the former indices, the historically used critical value of .10 or less (Medkser et al., 1994) has been challenged by Hu and Bentler (1999). They recommend a value of .08 or less. Per the logic above, our preference at this stage is to view the .08 as indicative of excellent fit, with the .10 acting as an upper limit.

In closing, there were three general reasons for selecting these four indices over many of the others. First, there has been consistent support for them in terms of their ability to distinguish between well-fitting and poor-fitting models and to do so accurately over a range of sample, study, and model characteristics. Second, the trade-offs they exhibit (e.g., the TLI's use in nested models, etc.) seem to be more aligned with the issues underlying measurement invariance (e.g., differences in factor loadings). Finally, they represent the range of classes into which fit indices have been categorized. The TLI and RNI are both incremental fit indices, but the former is a Type II incremental index, whereas the latter is a Type III (see Hu & Bentler, 1999). In contrast, both the RMSEA and the SRMR are referred to as absolute fit indices.

Differences between models. Important to inferences within the measurement invariance arena is examining the difference between a more restricted model (i.e., invariance constraints in place) and a less restricted model (i.e., a model in which those constraints are not in place). The tenability of the constraints is determined by whether the constrained model resulted in a significant worsening of fit relative to the less constrained model. The decision in practical terms is whether a more parsimonious model (i.e., the one with invariance constraints) fits the data just (or nearly) as well as a more complex model and thus can be used to explain the data.

The most frequently used tool for testing the difference between models is the chi-square difference test. The justification for doing so is typically attributed to Steiger, Shapiro, and Browne (1985) who demonstrated that incremental chi-square values are asymptotically independent test statistics. Recently, though, some researchers are questioning whether this should be the only indicator of difference between models (Brannick, 1995; Cheung & Rensvold, 1999c; Kelloway, 1995). The issue is that although researchers use a range of fit indices to determine overall model fit, they have only used the chi-square value to detect differences between models, creating a double standard. One attempt to address this was provided by Little (1997), who used a set of descriptive steps, each of which had to be met before he was satisfied that a constrained model was different from its less constrained baseline. The point is that Little had to rely on descriptive procedures for which there was little proof that a step would indeed result in a valid decision regarding the difference between models. Therein lie also the root problem and the reason why researchers appear to fall back on only a single standard to detect model differences. Specifically, whereas the chi-square has a known distribution from which we can specify critical values along different degrees of freedom, no known distributions exist for the other fit indices. Consequently, even if a researcher wanted to take the difference between TLI values of two nested models, for example, and from that infer whether one model fits less well than the other model, there is no standard against which to compare that difference value so that a valid decision will be made. Recent research has started to address this concern. In particular, Cheung and Rensvold (1999c) conducted a Monte Carlo simulation in which they examined differ-

ences in TLI, RMSEA, and comparative fit index (CFI) (Bentler, 1990) across all eight levels of measurement invariance and under varying study, model, and sample characteristics (e.g., large vs. small samples, different number of items per factor, etc.). They concluded that although examining differences in all three fit indices was superior to examining the difference in chi-square, they only reported the critical values for the change in CFI. Specifically, they claimed that changes in CFI of –.01 or less indicate that the invariance hypothesis should not be rejected, but when the differences lie between –.01 and –.02, the researcher should be suspicious that differences exist. Definite differences between models exist when the change in CFI is greater than –.02. No such critical levels were given for the TLI and RMSEA. Furthermore, no comparison was made between those critical values and the values of the difference in chi-square procedure.

Although in the appropriate direction, the Cheung and Rensvold (1999c) study is the first of its kind. Therefore, before we can have absolute confidence in their results, additional simulations need to be undertaken. In the meantime, we see no choice now other than to recommend that researchers remain with the chi-square difference procedure as the primary means for evaluating model differences. Based on Cheung and Rensvold's findings, though, we encourage researchers to also examine the differences in the practical fit indices and perhaps use the CFI criterion as a supplement.

Reference indicator selection. This issue arises if the researcher rejects the hypothesis of metric invariance (equal factor loadings) and undertakes tests for partial metric invariance. As reflected in Tables 1 and 2, this is not an infrequent practice, but what is not reflected here is the fact that this practice may not have been appropriately undertaken in the many cases. It is beyond the scope of this article to detail this issue, but Cheung and Rensvold (1999a) provide an excellent discussion of it. Partial metric invariance entails identifying and then freeing the "offending" invariance constraint or constraints that caused the model to fit poorly. *Offending* means that the item was calibrated to the underlying true score differently (calibrated to a different metric) in one or more groups. The tenability of freeing that constraint is tested by again taking the difference between this new, partial-metric invariance model and the baseline configural invariance model. If the new model now fits as well to the data as the baseline model, then the researcher purportedly knows which item or items caused the original metric invariance hypothesis to be untenable.

A potential problem in this procedure arises, though, when the researcher unknowingly selects the offending factor loading as the reference indicator. Recall that all factors require that one item be used as a referent indicator to assign a metric to the latent variable. If, however, the researcher unknowingly uses an item that in reality is the offending one that is responsible for the metric invariance hypothesis being rejected in the first place, then the partial metric invariance models will continue to fit poorly to the data. The poor fit, however, is not necessarily due to any "real" difference between the models. Rather, it is an artifact of having standardized the latent variable to different metrics as a function of using an item that itself has a different relationship to the latent variable from one group to the next. As noted earlier, Cheung and Rensvold (1999a) provide an excellent accounting of the issue, and they provide a heuristic (triangle heuristic) using a factor-ratio test that is intended to lower the probability of this problem arising. Again, though, their study represents the first of its kind, and as such,

additional research is required to further increase confidence that their procedure is viable.

An Empirical Example of Assessing Longitudinal ME/I

The following example is based on data reported earlier by Lance et al. (1999) in a study of organizational newcomers' adjustment to their employing organization. As argued repeatedly here, tests of ME/I should be routinely conducted prior to conducting tests aimed at evaluating cross-group differences. Lance et al. echoed this concern specifically as it relates to tests of ME/I as being a prerequisite to conducting LGM analysis of longitudinal change. As such, portions of the results reported here were presented earlier by Lance et al. as prerequisite analyses of ME/I prior to LGM analysis of change over time.

Participants and procedures. Complete data on variables described here were available from 104 newly hired employees of a large southeastern banking institution. Participants completed questionnaires containing (in addition to other measures) Vandenberg, Self, and Seo's (1994) modified versions of O'Reilly and Chatman's (1986) compliance (five items), internalization (five items), and identification (six items) organizational commitment scales. All items were anchored with 5-point Likert-type scales (1 = *strongly disagree* to 5 = *strongly agree*). Internal consistencies ranged between .70 and .93; Vandenberg et al. present additional details regarding these scales' development, evaluation, and psychometric properties.

Questionnaires were administered to participants on three separate occasions: (a) Time 1 (T1), at an orientation session in the newcomer participants' very first hour of employment in the organization; (b) Time 2 (T2), on the employees' 3-month employment anniversary; and (c) Time 3 (T3), on the employees' 6-month employment anniversary. Data reported here are from those participants who provided complete data from all three measurement waves. Thus, data for this example (a) relate to three constructs (compliance, internalization, and identification commitment), each of which is operationalized using multiple manifest indicators (i.e., items); (b) include no missing data; and (c) follow multiwave longitudinal structure with (approximately) equal measurement intervals (Willett & Sayer, 1994). A complete description of the sample, measures, and procedure is given in Lance et al. (1999).

Analyses. As we indicated earlier, analyses of time-structured data of the form reported here can be conducted either using a multisample approach or by inputting a single augmented covariance matrix for analysis. To take fuller advantage of the time-structured nature of the data, we opted for the latter approach so that the input covariance matrix was dimensioned as a *k*-item $\times$ 3-measurement occasion matrix for each organizational commitment dimension. However, doing so meant that our analytic approaches would differ from a more typical multisample analytic approach. To begin with, the first ME/I test in a multisample approach is an omnibus test of the equality of sample covariance matrices with, for example, "stacked" covariance matrices input to the LISREL program. However, this approach fails to take advantage of the time-structured nature of the sort of data reported here because it ignores longitudinal relationships between repeated measures on variables obtained on multiple occasions.

Rather, we conducted the omnibus test of equality of covariance matrices ($\Sigma^g = \Sigma^{g'}$) over time by (a) inputting the single *k*-item × 3-measurement occasion matrix for analysis for each organizational commitment dimension separately, (b) specifying that the number of latent variables equaled the number of observed variables on a single measurement occasion (i.e., the number of ξ variables equaled *K*, along with specifications that the factor pattern matrix was an identity matrix [i.e., $\Lambda = \mathbf{I}$] and that the matrix of residual variances and covariances was null [i.e., $\Theta_\delta = \mathbf{0}$]), (c) constraining variances of and covariances between like items/factors to be identical across measurement waves (i.e., T1, T2, and T3 item/factor variances and covariances were constrained to be equal within each occasion of measurement), and (d) constraining covariances across adjacent measurement waves to be equal (i.e., T1 – T2 item/factor covariances were constrained to be equal to T2 – T3 covariances, but these were not constrained to be equal to T1 – T3 covariances). In effect, this model tests not only for equality of within-occasion covariance matrices across measurement waves (as does the multisample approach) but also tests for homogeneity of first-order lagged associations across adjacent measurement waves. We refer to this model below as a test of invariant covariance matrices ($\Sigma^g = \Sigma^{g'}$) and as Model 0. The LISREL code for this parameterization is included in Appendix A. A well-fitting model of this form supports longitudinal measurement invariance, and further tests of specific aspects of ME/I are unnecessary; a poor-fitting model suggests that one or more specific forms of measurement inequivalence incurs over time and indicates that further tests of specific aspects of ME/I are warranted.

The next step (Model 1) was to test for configural invariance. We describe the model tested here in some detail because additional ME/I tests build on it. Recall that the test of configural invariance is a test of equivalent factor structures—in this case, across measurement occasions. Because each set of items was intended to reflect one (and only one) organizational commitment dimension (either compliance, internalization, or identification), the configural invariance model corresponded to a three-factor simple structure model in which all items were specified as loading on a single factor corresponding to the occasion in which they were measured. In effect, this model hypothesizes that a congeneric, unidimensional model holds in each measurement occasion. Additional specifications included the following: (a) For each time frame, one item's (the first item's) factor loading was fixed equal to 1.0 to set the scale of the factor corresponding to each measurement occasion; freely estimated factor loadings were not constrained to be equal across occasions. (b) The first item's intercept was fixed equal to 0.0 to set the mean of each factor. (c) Factor variances, covariances, and means were freely estimated and allowed to be heterogeneous across measurement occasions. (d) Unique variances were freely estimated and allowed to be heterogeneous across occasions. Also, covariances between like items' uniquenesses were estimated across occasions (see Vandenberg & Self, 1993). The rationale for estimating covariances between like items' uniquenesses over time follows from the ideas that (a) item variances are composed of both common and unique components (i.e., $\sigma^2_{Xk} = \lambda^2_k \Phi + \Theta_{\delta k}$ in a unidimensional congeneric model), (b) that unique variance is composed of both item-specific variance and nonsystematic measurement error (i.e., $\Theta_{\delta k} = s^2_k + e^2_k$), and (c) item-specific components may be temporally stable. Thus,

estimating covariances among like items' uniquenesses controls for temporal stability in item specificities. A well-fitting configural invariance model (Model 1) suggests that a unidimensional congeneric measurement model is plausible across all measurement occasions and that additional ME/I tests may proceed. A poor-fitting configural invariance model suggests that even "weak factorial invariance" (Horn & McArdle, 1992) does not incur and therefore that additional ME/I tests are precluded.

The next ME/I test (Model 2) was of metric invariance, in which like items' factor loadings were constrained to be equal across measurement occasion. Thus, the additional constraint that $\Lambda^{g} = \Lambda^{g'}$ was imposed on Model 1 to define Model 2. (At least partial) metric invariance must be tenable before proceeding to additional ME/I tests. The next ME/I model (Model 3) invoked additional equality constraints on like items' intercepts across occasions. As such, Model 3 represents the scalar invariance model. Subsequently, and in the following order, we invoked the remaining constraints across measurement occasions: (a) Model 4, invariant uniquenesses ($\Theta^{g} = \Theta^{g'}$); (b) Model 5, invariant factor variances ($\Phi^{g}_{j} = \Phi^{g'}_{j}$); (c) Model 6, invariant factor covariances ($\Phi^{g}_{jj'} = \Phi^{g'}_{jj'}$); and (d) Model 7, invariant factor means ($\kappa^{g} = \kappa^{g'}$). We used the LISREL 8.12 program (Jöreskog & Sörbom, 1996) to conduct all analyses. Appendix B contains the LISREL code for Model 7 for the compliance items.

Results. Table 3 shows results of ME/I tests for compliance commitment. Results for Model 0 (invariant covariance matrices) indicated excellent model fit. The model $\chi^2(65) = 66.69$ is statistically nonsignificant, and every other fit index (except for the RNI) also supports the idea that Model 0 fits the augmented longitudinal covariance matrix very well. Consequently, longitudinally invariant measurement operations are indicated, and further ME/I tests are not necessary and need not be undertaken under normal circumstances. Nevertheless, we report the remaining ME/I test results for illustrative purposes only. By all indications, both the configural invariance model (Model 1) and the metric invariance model (Model 2) also provide excellent fits to the data. The nonsignificant difference χ^2 ($\Delta\chi^2$) and the very small change in CFI (ΔCFI) (Cheung & Rensvold, 1999c) between Models 1 and 2 give additional support to the idea that the invariance constraints imposed by Model 2 (i.e., $\Lambda^{g} = \Lambda^{g'}$) did not significantly worsen model fit as compared to Model 1, thus supporting the viability of these constraints. In fact, the overall goodness-of-fit indices, as well as tests of differences in fit between adjacent (nested) models reported in Table 3, support inferences of longitudinal measurement invariance with respect to every aspect of ME/I for compliance commitment. Thus, compliance demonstrated both measurement and structural invariance longitudinally. As such, results corresponding to specific ME/I illustrate why it is unnecessary to test for specific aspects of ME/I when the initial omnibus test of equivalent covariance matrices (i.e., Model 0) indicates satisfactory ME/I.

However, results for identification commitment, shown in Table 4, paint quite a different picture than that for compliance. By all standards, Model 0 provided a poor fit to the data: Model 0 was rejected statistically ($\chi^2[91] = 186.78$, $p < .01$) and on the basis of criterion values recommended for the remaining fit indices reported in Table 4 (i.e., SRMSR > .08; RMSEA > .06; TLI, RNI, and CFI < .95). This suggests that further ME/I tests are warranted to identify the source(s) of longitudinal measurement inequivalence. The first of these tests, Model 1 (configural invariance) provided a poor

Table 3
Tests of Longitudinal measurement Equivalence for Organizational Commitment Dimensions: Compliance

Model	df	χ^2	*SRMSR*	*RMSEA*	*TLI*	*RNI*	*CFI*	Δdf	$\Delta\chi^2$	ΔCFI
0. Invariant covariance matrices ($\Sigma^g = \Sigma^{g'}$)	65	66.69	.062	.000	.99	.90	1.00	—	—	—
1. Configural invariance	72	81.61	.058	.025	.97	.90	.98	—	—	—
1 versus 2	—	—	—	—	—	—	—	8	3.30	.01
2. Metric invariance ($\Lambda^g = \Lambda^{g'}$)	80	84.91	.062	.020	.98	.93	.99	—	—	—
2 versus 3	—	—	—	—	—	—	—	8	13.85	–.02
3. Scalar invariance ($\tau^g = \tau^{g'}$)	88	98.76	.062	.014	.97	.94	.97	—	—	—
3 versus 4	—	—	—	—	—	—	—	10	7.85	.01
4. Invariant uniquenesses ($\Theta^g = \Theta^{g'}$)	98	106.61	.065	.004	.98	.97	.98	—	—	—
4 versus 5	—	—	—	—	—	—	—	2	2.88	.00
5. Invariant factor variances ($\Phi_j^g = \Phi_j^{g'}$)	100	109.49	.077	.014	.98	.97	.98	—	—	—
5 versus 6	—	—	—	—	—	—	—	2	0.54	.00
6. Invariant factor covariances ($\Phi_{ij}^g = \Phi_{ij}^{g'}$)	102	110.03	.079	.007	.98	.98	.98	—	—	—
6 versus 7	—	—	—	—	—	—	—	2	0.89	.00
7. Invariant factor means ($\kappa^g = \kappa^{g'}$)	104	110.92	.079	.004	.98	.99	.98	—	—	—

NOTE: SRMSR = standardized root mean squared residual; RMSEA = root mean squared error of approximation; TLI = Tucker-Lewis index; CFI = comparative fit index.

Table 4
Tests of Longitudinal measurement Equivalence for Organizational Commitment Dimensions: Identification

Model	df	χ^2	*SRMSR*	*RMSEA*	*TLI*	*RNI*	*CFI*	Δdf	$\Delta\chi^2$	ΔCFI
0. Invariant covariance matrices ($\Sigma^g = \Sigma^{g'}$)	91	186.78*	.150	.092	.87	.84	.92	—	—	—
1. Configural invariance	114	199.60*	.067	.074	.90	.91	.93	—	—	—
1 versus 2	—	—	—	—	—	—	—	10	9.69	.00
2. Metric invariance ($\Lambda^g = \Lambda^{g'}$)	124	209.29*	.073	.072	.91	.94	.93	—	—	—
2 versus 3	—	—	—	—	—	—	—	10	31.54*	–.02
3. Scalar invariance ($\tau^g = \tau^{g'}$)	134	240.83*	.081	.079	.90	.93	.91	—	—	—
3 versus 4	—	—	—	—	—	—	—	12	47.09*	–.03
4. Invariant uniquenesses ($\Theta^g = \Theta^{g'}$)	146	287.92*	.085	.091	.88	.93	.88	—	—	—
4 versus 5	—	—	—	—	—	—	—	2	15.49*	–.01
5. Invariant factor variances ($\Phi_j^g = \Phi_j^{g'}$)	148	303.41*	.150	.095	.87	.92	.87	—	—	—
5 versus 6	—	—	—	—	—	—	—	2	23.00*	–.02
6. Invariant factor covariances ($\Phi_{ij}^g = \Phi_{ij}^{g'}$)	150	326.41*	.170	.100	.85	.90	.85	—	—	—
6 versus 7	—	—	—	—	—	—	—	2	24.06*	.01
7. Invariant factor means ($\kappa^g = \kappa^{g'}$)	152	350.47*	.170	.110	.83	.89	.84	—	—	—

NOTE: SRMSR = standardized root mean squared residual; RMSEA = root mean squared error of approximation; TLI = Tucker-Lewis index; CFI = comparative fit index.
*p < .01.

However, results for identification commitment, shown in Table 4, paint quite a different picture than that for compliance. By all standards, Model 0 provided a poor fit to the data: Model 0 was rejected statistically ($\chi^2[91] = 186.78$, $p < .01$) and on the basis of criterion values recommended for the remaining fit indices reported in Table 4 (i.e., SRMSR > .08; RMSEA > .06; TLI, RNI, and CFI < .95). This suggests that further ME/I tests are warranted to identify the source(s) of longitudinal measurement inequivalence. The first of these tests, Model 1 (configural invariance) provided a poor fit to the data: Model 1 was rejected statistically ($\chi^2[114] = 199.60$, $p < .01$) and on the basis of all other overall fit indices. Thus, the identification commitment's scale items functioned nonequivalently across measurement occasions. To explore why, we conducted exploratory principal components analyses (PCAs) on the identification items separately within each measurement occasion. Although the PCAs suggested that the items formed a unidimensional scale at each measurement wave, the first eigenvalues indicated that identification was an increasingly unified construct over time (eigenvalues were 3.26 [54% of the total variance], 3.52 [59%], and 4.45 [74%] for T1 through T3, respectively).

More important, failure to support configural invariance militates against meaningful tests of additional aspects of ME/I. Nevertheless, we present results of these additional tests for illustration, keeping in mind that substantive inferences from these tests are limited at best and may not be meaningful at all. Were these remaining tests' results interpretable, they would suggest measurement inequivalence with respect to each aspect of ME/I, save metric invariance, according to the $\Delta\chi^2$ criterion; conclusions based on ΔCFI (i.e., $\Delta CFI \leq -.02$; Cheung & Rensvold, 1999c) were far more ambiguous. Thus, although we appreciate the logic and spirit within which the ΔCFI was developed, we recommend caution in its application pending further research and evaluation.

Finally, Table 5 shows ME/I test results for internalization commitment. As for identification commitment, results for Model 0 indicated some form of measurement inequivalence for internalization ($\chi^2[65] = 136.92$, $p < .01$; SRMSR > .08; RMSEA > .06; TLI, RNI, and CFI < .95). Unlike identification, however, the test for configural invariance (Model 1) indicated acceptable model fit by every standard except the RNI, suggesting that some other specific aspect of ME/I was responsible for longitudinal measurement inequivalence. Further tests supported the viability of the metric invariance hypothesis (Model 1 vs. Model 2 $\Delta\chi^2[8] = 5.33$, $p > .01$; $\Delta CFI = .00$) but not the hypothesis of scalar invariance (Model 2 vs. Model 3 $\Delta\chi^2[8] = 41.65$, $p < .01$; $\Delta CFI = -.04$). Results also suggested that the factor means were not invariant (Model 6 vs. Model 7 $\Delta\chi^2[2] = 58.56$, $p < .01$; $\Delta CFI = -.07$). We explored these results in more detail by investigating possible partial ME/I.

The first three rows of Table 6 are reproduced from corresponding rows in Table 5 and support inferences that internalization items demonstrated both configural and metric invariance longitudinally. A close examination of Model 2's results indicated that internalization items' T1 intercepts were uniformly and significantly higher than were T2's and T3's, suggesting the source of scalar inequivalence indicated earlier in Table 5. Model 3 in Table 6 tests for partial scalar invariance, that is, $T1 \neq T2 = T3$ intercepts.[2] Table 6 shows that invoking only partial scalar invariance (PSI) resulted in generally acceptable model fit. Although the $\Delta\chi^2$ was statistically significant (Model 2 vs. Model 3 $\Delta\chi^2[5] = 19.00$, $p < .01$), Model 3 demonstrated good fit according to the overall χ^2 statistic ($\chi^2[85] = 105.90$, $p > .01$) and all other goodness-of-fit indices. However,

Table 5
Tests of Longitudinal Measurement Equivalence for Organizational Commitment Dimensions: Internalization

Model	df	χ^2	*SRMSR*	*RMSEA*	*TLI*	*RNI*	*CFI*	Δdf	$\Delta\chi^2$	ΔCFI
0. Invariant covariance matrices ($\Sigma^g = \Sigma^{g'}$)	65	136.92*	.110	.100	.87	.83	.92	—	—	—
1. Configural invariance	72	81.57	.047	.029	.98	.95	.99	—	—	—
1 versus 2	—	—	—	—	—	—	—	8	5.33	.00
2. Metric invariance ($\Lambda^g = \Lambda^{g'}$)	80	86.90	.056	.021	.99	.96	.99	—	—	—
2 versus 3	—	—	—	—	—	—	—	8	41.65*	–.04
3. Scalar invariance ($\tau^g = \tau^{g'}$)	88	128.55*	.064	.055	.95	.93	.95	—	—	—
3 versus 4	—	—	—	—	—	—	—	10	22.93	–.01
4. Invariant uniquenesses ($\Theta^g = \Theta^{g'}$)	98	151.48*	.067	.062	.94	.93	.94	—	—	—
4 versus 5	—	—	—	—	—	—	—	2	2.60	.00
5. Invariant factor variances ($\Phi_j^g = \Phi_j^{g'}$)	100	154.08*	.086	.062	.94	.94	.94	—	—	—
5 versus 6	—	—	—	—	—	—	—	2	4.40	.00
6. Invariant factor covariances ($\Phi_{ij}^g = \Phi_{ij}^{g'}$)	102	158.48*	.100	.063	.94	.93	.94	—	—	—
6 versus 7	—	—	—	—	—	—	—	2	58.56*	–.07
7. Invariant factor means ($\kappa^g = \kappa^{g'}$)	104	217.04*	.110	.100	.87	.87	.87	—	—	—

NOTE: SRMSR = standardized root mean squared residual; RMSEA = root mean squared error of approximation; TLI = Tucker-Lewis index; CFI = comparative fit index.
*p < .01.

Table 6
Tests of Partial Invariance Constraints: Internalization

Model	df	χ^2	*SRMSR*	*RMSEA*	*TLI*	*RNI*	*CFI*	Δdf	$\Delta\chi^2$	ΔCFI
1. Configural invariance	72	81.57	.047	.029	.98	.95	.99	—	—	—
1 versus 2	—	—	—	—	—	—	—	8	5.33	.00
2. Metric invariance	80	86.90	.056	.021	.99	.96	.99	—	—	—
2 versus 3	—	—	—	—	—	—	—	5	19.00*	–.01
3. Metric invariance and partial scalar invariance (PSI)	85	105.90	.054	.041	.97	.95	.98	—	—	—
3 versus 4	—	—	—	—	—	—	—	10	25.00*	–.02
4. Invariant uniquenesses and PSI	95	130.90*	.057	.055	.96	.95	.96	—	—	—
3 versus 4′	—	—	—	—	—	—	—	5	10.40	–.01
4′. Partially invariant uniquenesses (PIU) and PSI	90	116.30	.057	.044	.97	.96	.97	—	—	—
4′ versus 5	—	—	—	—	—	—	—	2	6.66	.00
5. PSI, PIU, and invariant factor variances	92	122.96	.080	.045	.97	.97	.97	—	—	—
5 versus 6	—	—	—	—	—	—	—	2	0.65	.00
6. Model 5 plus invariant factor covariances	94	123.64	.095	.046	.96	.97	.97	—	—	—
6 versus 7	—	—	—	—	—	—	—	2	9.35*	–.01
7. Model 6 plus invariant factor means	96	132.99*	.095	.031	.95	.98	.96	—	—	—

NOTE: SRMSR = standardized root mean squared residual; RMSEA = root mean squared error of approximation; TLI = Tucker-Lewis index; CFI = comparative fit index.
*$p < .01$

when uniquenesses were constrained to be invariant (in addition to maintaining metric and partial scalar invariance constraints [Model 4]), model fit worsened significantly (Model 4 $\chi^2[95] = 130.90$, $p < .01$; Model 3 vs. Model 4 $\Delta\chi^2[10] = 25.00$, $p < .01$; ΔCFI = –.02). This was because T3 uniquenesses were more heterogeneous (both higher and lower) as compared to T1 and T2. Controlling for PSI (T1 ≠ T2 = T3 intercepts) and partially invariant uniquenesses (T1 = T2 ≠ T3 unique variances) resulted in acceptable model fit overall (Model 4′ $\chi^2[90] = 116.30$, $p > .01$; SRMSR < .08; RMSEA < .06; TLI, RNI, and CFI > .95) and in comparison to the model that invoked only PSI (Model 3 vs. Model 4′ $\Delta\chi^2[5] = 10.40$, $p > .01$). The remainder of Table 6 shows that tests for structural invariance with respect to latent variances and covariances were viable but that there were mean differences in internalization over time.

Summary. The preceding example illustrates several points. First, in the case of compliance, it illustrates how researchers can and should more routinely assess and demonstrate cross-group ME/I prior to conducting cross-group comparisons. As we noted earlier, demonstration of acceptable traditional psychometric properties (e.g., internal consistency, predictive validity) is important but not sufficient in determining whether psychological measures function equivalently across groups. Second, in the case of identification, the example illustrates how tests of ME/I can be used to detect nonequivalent measurement operations across groups that might likely be otherwise undetected using more traditional psychometric scale evaluation approaches. Knowing that a scale functions nonequivalently across intended comparison groups should, at a minimum, mandate caution on the part of researchers intending to compare groups based on the nonequivalent measure. More generally, comparisons on nonequivalent measures should not be undertaken. Third, in the case of internalization, the example shows how some (relatively minor) sources of measurement inequivalence can be controlled for by invoking partial ME/I constraints. Fourth, the example illustrates how overall (χ^2, SRMSR, RMSEA, TLI, RNI, and CFI) and incremental ($\Delta\chi^2$, ΔCFI) goodness-of-fit indices can be combined to make decisions regarding overall model fit and the plausibility of various ME/I constraints. Relatedly, we note the comparative usefulness of the (traditional) $\Delta\chi^2$ statistic and a (recently proposed, Cheung & Rensvold, 1999c) ΔCFI in evaluating the viability of ME/I constraints in nested model comparisons. Pending additional research on incremental fit indices such as the ΔCFI, we recommend continued use of the $\Delta\chi^2$ statistic. Finally, the example demonstrates the sequence of ME/I tests discussed repeatedly in this article in use to evaluate ME/I issues for three scales that demonstrated different degrees of measurement equivalence/inequivalence. Figure 2 presents an overview of the logical sequencing of these tests.

Recommendations and Conclusions

Figure 2 presents a flowchart detailing our recommended sequencing of ME/I tests and some of the critical decision points along that sequence. Our recommended sequencing differs in two broad respects from past treatments of the sequence issue (e.g., Bagozzi & Edwards, 1998; Byrne et al., 1989; Jöreskog, 1974; Schmitt, 1982; for another flowchart, see Steenkamp & Baumgartner, 1998). First, we felt that some depictions of the partial invariance issue at the various steps have presented the issue

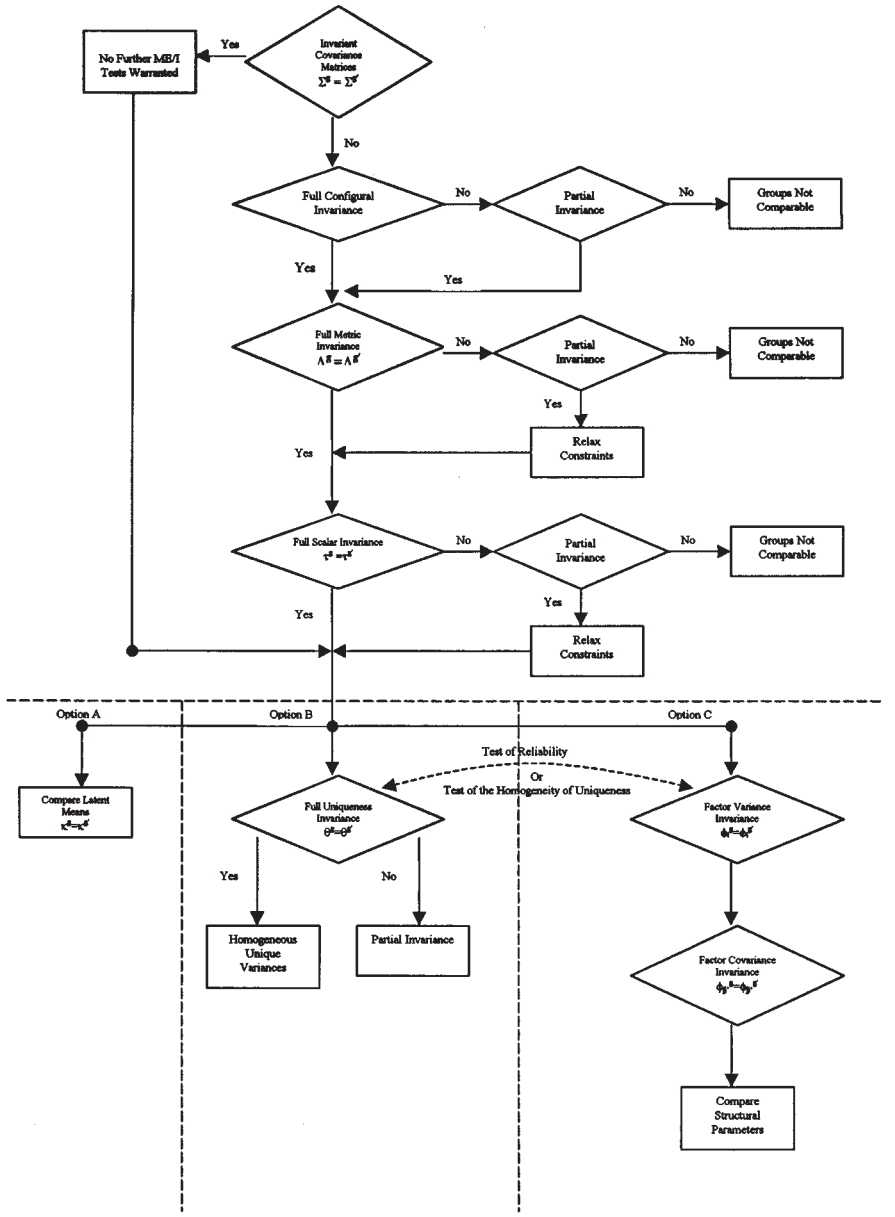

Figure 2: A Flowchart of the Recommended Sequence for Conducting Tests of Measurement Equivalence/Invariance

too liberally (e.g., Byrne et al., 1989; Steenkamp & Baumgartner, 1998). The message from the reviewed studies could be interpreted by the uninitiated as one of "repeat it as often as needed until the problem disappears." In contrast, we present partial invariance as a critical decision point when the researcher must carefully consider whether to engage in the practice in the first place. Doing so requires beginning with a strong

theoretical justification as to which constraints are relaxed or which items a researcher may consider eliminating from the analysis. In addition to theory, the researcher should not rely solely on the modification index (the only index advocated by some researchers) (Byrne et al., 1989; Steenkamp & Baumgartner, 1998) but should also examine other properties such as the degree of expected parameter change or the absolute differences in factor loading values of like items between groups. As a critical decision point, the researcher may feel that there is no justification for engaging in partial invariance and thus must conclude that the groups are not comparable. At other times, there may be adequate justification for engaging in the practice, and thus the researcher relaxes a reasonable number of constraints and continues in the sequence of tests. It is important to note that by placing partial invariance as an element to examining configural invariance, we are not advocating altering the conceptual domain of the measures. Rather, we are stating that justification may exist under some circumstances to proceed with further tests when configural invariance was statistically untenable at first, but the latter was due to a failure to support equivalence of all constructs in a multiple construct application. Returning to the example presented earlier, Lance et al. (1999) proceeded with LGM analysis of change only with respect to compliance and internalization commitment only because there were sufficient indications that identification did not satisfy the requirement of the configural invariance prerequisite to analysis of longitudinal change. Thus, Lance et al. (1999) were still able to evaluate their conceptual model of the development of newcomer organizational commitment but with respect to only two of the three forms of commitment conceptualized initially.

The second difference between the current flowchart and previous depictions of the ME/I sequence is our treatment of the consequences or results. At a minimum, we felt there were three options. Option A is a comparison of latent means after completing the test for scalar invariance. Thus, if substantive reasoning calls for a group comparison at the mean level, one could proceed at this point without having to impose the other constraints implied under Options B and C. Latent means are theoretically perfectly reliable, and thus testing the invariance of uniqueness terms before testing latent mean differences gains nothing. Similarly, latent means are independent of the factor variances and covariances. Thus, again, testing the invariance of the factor variances and covariances gains nothing before testing the comparability of the latent means. Option B may occur under circumstances when the researcher is concerned about the homogeneity of unique variances. This may be of interest, for example, in determining whether elders respond to surveys less reliably with progressive dementia symptoms. Perhaps our most radical departure from past practices is the recommendation represented in Option C. Because factor variances and covariances are themselves structural parameters, it is unreasonable to examine their invariance properties until the invariance of the parameters of the measurement model has been established (see Anderson & Gerbing, 1988, as to why the measurement model should be established first before examining structural parameters). Our view treats the test of factor variance and covariance invariance as omnibus tests. Specifically, assume that the substantive goal of the researcher is to test whether a causal model among the set of latent variables fits equally well in the comparison groups. Testing the variance-covariance matrix of the underlying factors is analogous to testing the covariance matrices in Step 1 of Figure 2. If the null hypotheses are not rejected, then a structural model

imposed among the latent variables will be invariant between groups. On the other hand, if the null hypothesis is untenable, then one can expect that at least some differences likely exist between groups as to how the latent variables relate to one another (e.g., Vandenberg & Scarpello, 1990).

To close, two other aspects of Figure 2 need to be noted. The first is the arrow from the "no further ME/I tests warranted" box to the intersection just under the "full scalar invariance" point. If "no further ME/I tests" are warranted, the researcher may simply proceed with a comparison of the groups (assuming substantive reasons exist to do so) using traditional statistical comparison procedures such as ANOVA or *t* test. What we wished to indicate with the arrow, however, is that other options exist as well that may be directly derived from a SEM approach. For example, rather than switch to traditional comparative procedures, the researcher could complete the same group comparison but do so at the latent mean level (Option A). Or Option C would permit the researcher to test differences in model parameters (comparable to regression coefficients) assuming the parameters may vary as a function of some group characteristic (e.g., Vandenberg & Scarpello, 1990). Our point simply is that the researcher may stay within the same analytical procedure without having to switch from this SEM approach to some other approach.

The second aspect to highlight is represented by the dashed arrow between "full uniqueness invariance" and "factor variance invariance." This serves as a reminder that whether full uniqueness invariance is considered a test of reliability or a test of homogeneity of uniqueness terms depends on the test of factor variance invariance. If the latter test fails to reject the null hypothesis, then and only then is it a test of reliability. Tests of ME/I have historically not been conducted as often as need be in organizational research. The intent of Figure 2 is to facilitate future applications of tests of ME/I.

APPENDIX A
LISREL Code for Omnibus Test of Equality of Covariance Matrices: Compliance Commitment

```
OMNIBUS TEST OF EQUALITY OF COVARIANCE MATRICES: COMPLIANCE
DA NI=48 NG=1 NO=104 MA=CM
LA FI=A:\LABELS.TXT
RA FI=A:\DATA.DAT

***COMMENT
***The following statements select out from the total data set the five COMPLIANCE
***items corresponding to T1, T2, and T3, respectively
***COMMENT

SE
*
12 13 14 15 16 28 29 30 31 32 44 45 46 47 48 /
MO NX=15 NK=15 LX=DI,FI TD=ZE TX=FR KA=FI PH=SY,FR
```

```
***COMMENT
***The following pattern matrix imposes equality constraints on estimates for
***like-numbered parameters.
***COMMENT

PA PH
*
99
2 98
3 4 97
5 6 7 96
8 9 10 11 95
12 13 14 15 16 99
17 18 19 20 21 2 98
22 23 24 25 26 3 4 97
27 28 29 30 31 5 6 7 96
32 33 34 35 36 8 9 10 11 95
37 38 39 40 41 12 13 14 15 16 99
42 43 44 45 46 17 18 19 20 21 2 98
47 48 49 50 51 22 23 24 25 26 3 4 97
52 53 54 55 56 27 28 29 30 31 5 6 7 96
57 58 59 60 61 32 33 34 35 36 8 9 10 11 95
***COMMENT
***The following statements (in addition to the specification that LX=DI,FI
***in the MO line statement ensure that the factor pattern matrix is fixed
***equal to an identity matrix.
***COMMENT

VA 1.0 LX(1,1) LX(2,2) LX(3,3) LX(4,4) LX(5,5)
VA 1.0 LX(6,6) LX(7,7) LX(8,8) LX(9,9) LX(10,10)
VA 1.0 LX(11,11) LX(12,12) LX(13,13) LX(14,14) LX(15,15)
***COMMENT

***The following statements constrain items' intercepts to be identical across
***measurement waves.
***COMMENT

EQ TX 1 TX 6 TX 11
EQ TX 2 TX 7 TX 12
EQ TX 3 TX 8 TX 13
EQ TX 4 TX 9 TX 14
EQ TX 5 TX 10 TX 15
OU AD=OFF SS SC SE TV IT=600
```

APPENDIX B
LISREL Code for Test of Invariant Factor Means: Compliance Commitment

```
TEST OF EQUAL FACTOR MEANS: COMPLIANCE
DA NI=48 NG=1 NO=104 MA=CM
LA FI=A:\LABELS.TXT
RA FI=A:\DATA.DAT

SE
*
12 13 14 15 16 28 29 30 31 32 44 45 46 47 48 /
MO NX=15 NK=3 LX=FR TD=SY,FR TX=FR KA=FR PH=SY,FR

***COMMENT
***The following lines define labels for the latent compliance variables at T1,
***T2, and T3, respectively
***COMMENT
LK
COMPLI1 COMPLI2 COMPLI3
PA LX
5(1 0 0) 5(0 1 0) 5(0 0 1)

***COMMENT
***The following lines fix the first item's factor loading equal to 1.0 within each
***measurement occasion
***COMMENT

FI LX 1 1 LX 6 2 LX 11 3
VA 1 LX 1 1 LX 6 2 LX 11 3

***COMMENT
***The following lines fix the first item's intercept equal to 0.0 within each
***measurement occasion
***COMMENT

FI TX 1
VA 0 TX 1
FI TX 6
VA 0 TX 6
FI TX 11
VA 0 TX 11

***COMMENT
***The following lines invoke invariance constraints for like items' intercepts
***across measurement occasions.
***COMMENT
```

```
EQ TX 2 TX 7 TX 12
EQ TX 3 TX 8 TX 13
EQ TX 4 TX 9 TX 14
EQ TX 5 TX 10 TX 15

***COMMENT
***The following lines invoke invariance constraints for like items' factor loadings
***across measurement occasions.
***COMMENT

EQ LX 2 1 LX 7 2 LX 12 3
EQ LX 3 1 LX 8 2 LX 13 3
EQ LX 4 1 LX 9 2 LX 14 3
EQ LX 5 1 LX 10 2 LX 15 3
PA PH
*
1
1 1
1 1 1

***COMMENT
***The following pattern matrix includes covariances among like items' uniquenesses
***across measurement occasions.
***COMMENT

PA TD
*
1
0 1
0 0 1
0 0 0 1
0 0 0 0 1
1 0 0 0 0 1
0 1 0 0 0 0 1
0 0 1 0 0 0 0 1
0 0 0 1 0 0 0 0 1
0 0 0 0 1 0 0 0 0 1
1 0 0 0 0 1 0 0 0 0 1
0 1 0 0 0 0 1 0 0 0 0 1
0 0 1 0 0 0 0 1 0 0 0 0 1
0 0 0 1 0 0 0 0 1 0 0 0 0 1
0 0 0 0 1 0 0 0 0 1 0 0 0 0 1

***COMMENT
```

```
***The following lines invoke invariance constraints for like items' uniquenesses
***across measurement occasions.
***COMMENT

EQ TD 1 1 TD 6 6 TD 11 11
EQ TD 2 2 TD 7 7 TD 12 12
EQ TD 3 3 TD 8 8 TD 13 13
EQ TD 4 4 TD 9 9 TD 14 14
EQ TD 5 5 TD 10 10 TD 15 15
***COMMENT
***The following lines invoke invariance constraints for factor variances and
covariances
***across measurement occasions.
***COMMENT

EQ PH 1 1 PH 2 2 PH 3 3
EQ PH 2 1 PH 3 1 PH 3 2

***COMMENT
***The following line invokes invariance constraints for factor means
***across measurement occasions.
***COMMENT

EQ KA 1 KA 2 KA 3
ST .5 ALL
OU AD=OFF SS SC SE TV IT=600
```

Notes

1. Or, more technically, sets of classes of events; see James, Mulaik, and Brett (1982).
2. With the constraints remaining that the first item's intercept was fixed equal to zero within each time period and that equality constraints were retained for the second item. These remaining constraints are necessary for model identification (see Steenkamp & Baumgartner, 1998).

References

Alwin, D. F., & Jackson, D. J. (1981). Applications of simultaneous factor analysis to issues of factorial invariance. In D. J. Jackson & E. F. Borgatta (Eds.), *Factor analysis and measurement in sociological research* (pp. 249-279). Beverly Hills, CA: Sage.

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

Armenakis, A. A., & Zmud, R. W. (1979). Interpreting the measurement of change in organizational research. *Personnel Psychology*, *32*, 709-723.

Babcock, R. L., Laguna, K. D., & Roesch, S. C. (1997). A comparison of the factor structure of processing speed for younger and older adults: Testing the assumption of measurement equivalence across age groups. *Psychology of Aging*, *12*, 268-276.

Babyak, M. A., Snyder, C. R., & Yoshinobu, L. (1993). Psychometric properties of the Hope Scale: A confirmatory factor analysis. *Journal of Research in Personality*, *27*, 154-169.

Bagozzi, R. P., & Edwards, J. R. (1998). A general approach for representing constructs in organizational research. *Organizational Research Methods*, *1*, 45-87.

Bagozzi, R. P., & Heatherton, T. F. (1994). A general approach to representing multifaceted personality constructs: Application to state self-esteem. *Structural Equation Modeling*, *1*, 35-67.

Bagozzi, R. P., & Phillips, L. W. (1991). Assessing construct validity in organizational research. *Administrative Science Quarterly*, *36*, 421-458.

Bartunek, J. M., & Franzak, F. J. (1988). The effects of organizational restructuring on frames of reference and cooperation. *Journal of Management*, *14*, 579-592.

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

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

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

Bollen, K., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin*, *110*, 305-314.

Bollen, K. A., & Long, J. S. (1993). *Testing structural equation models*. Newbury Park, CA: Sage.

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

Brennan, R. L. (1983). *Elements of generalizability theory*. Iowa City, IA: American College Testing.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. *Sociological Methods and Research*, *21*, 230-258.

Burke, M. J., Brief, A. P., George, J. M., Roberson, L., & Webster, J. (1989). Measuring affect at work: Confirmatory analyses of competing mood structures with conceptual linkage to cortical regulatory systems. *Journal of Personality and Social Psychology*, *57*, 1091-1102.

Byrne, B. M. (1988). Measuring adolescent self-concept: Factorial validity and equivalency of the SDQ III across gender. *Multivariate Behavioral Research*, *23*, 361-375.

Byrne, B. (1989). *A primer of LISREL: Basic applications and programming for confirmatory factor analytic models*. New York: Springer Verlag.

Byrne, B. M. (1991). The Maslach Burnout Inventory: Validating factorial structure and invariance across intermediate, secondary, and university educators. *Multivariate Behavioral Research*, *26*, 583-605.

Byrne, B. M. (1994). Testing for the factorial validity, replication, and invariance of a measuring instrument: A paradigmatic application based on the Maslach Burnout Inventory. *Multivariate Behavioral Research*, *29*, 289-311.

Byrne, B. M., & Baron, P. (1993). The Beck Depression Inventory: Testing and cross-validating a hierarchical factor structure for nonclinical adolescents. *Measurement and Evaluation in Counseling and Development*, *26*, 164-178.

Byrne, B. M., & Baron, P. (1994). Measuring adolescent depression: Tests of equivalent factorial structure for English and French versions of the Beck Depression Inventory. *Applied Psychology: An International Review*, *43*, 33-47.

Byrne, B. M., Baron, P., & Campbell, T. L. (1993). Measuring adolescent depression: Factorial validity and invariance of the Beck Depression Inventory across gender. *Journal of Research on Adolescence*, *3*, 127-143.

Byrne, B. M., Baron, P., & Campbell, T. L. (1994). The Beck Depression Inventory (French version): Testing for gender-invariant factorial structure for nonclinical adolescents. *Journal of Adolescent Research*, *9*, 166-179.

Byrne, B. M., & Shavelson, R. J. (1987). Adolescent self-concept: Testing the assumption of equivalent structure across gender. *American Educational Research Journal*, *24*, 365-385.

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

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

Campion, M. A. (1993). Article review checklist: A criterion checklist for reviewing research articles in applied psychology. *Personnel Psychology*, *46*, 705-718.

Chan, D. (1996). Cognitive misfit of problem-solving style at work: A facet of person-organization fit. *Organizational Behavior and Human Decision Processes*, *68*, 194-207.

Chan, D. (1997). Racial subgroup differences in predictive validity perceptions on personality and cognitive ability tests. *Journal of Applied Psychology, 82*, 311-320.

Chan, D. (1998). The conceptualization and analysis of change over time: An integrative approach incorporating longitudinal mean and covariance structures analysis (LMACS) and multiple indicator latent growth modeling (MLGM). *Organizational Research Methods*, *1*, 421-483.

Chan, D., & Schmitt, N. (1997). Video-based versus paper-and-pencil method of assessment in situational judgment tests: Subgroup differences in test performance and face validity perceptions. *Journal of Applied Psychology, 42*, 143-159.

Chan, D., Schmitt, N., Sacco, J. M., & DeShon, R. P. (1998). Understanding pretest and posttest reactions to cognitive ability and personality tests. *Journal of Applied Psychology*, *83*, 471-485.

Cheung, G. W., & Rensvold, R. B. (1999a, August). *The effects of model parsimony and sampling error on the fit of structural equation models*. Paper presented at the 1999 conference of the Academy of Management, Chicago.

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

Cheung, G. W., & Rensvold, R. B. (1999c, August). *What constitutes significant differences in evaluating measurement invariance?* Paper presented at the 1999 conference of the Academy of Management, Chicago.

Cole, D. A., Gondoli, D. M., & Peeke, L. G. (1998). Structure and validity of parent and teacher perceptions of children's competence: A multitrait-multimethod-multigroup investigation. *Psychological Assessment*, *10*, 241-249.

Cole, D. A., & Maxwell, S. E. (1985). Multitrait-multimethod comparisons across populations: A confirmatory factor analytic approach. *Multivariate Behavioral Research*, *20*, 389-417.

Cole, D. A., Maxwell, S. E., Arvey, R., & Salas, E. (1993). Multivariate group comparisons of variable systems: MANOVA and structural equation modeling. *Psychological Bulletin*, *114*, 174-184.

Collins, J. M., & Gleaves, D. H. (1998). Race, job applicants, and the five-factor model of personality: Implications for Black psychology, industrial/organizational psychology, and the five-factor theory. *Journal of Applied Psychology, 83*, 531-544.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Fort Worth, TX: Harcourt Brace.

Cronbach, L. J., & Furby, L. (1971). How should we measure "change"—Or should we? *Psychological Bulletin*, *74*, 68-80.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability scores and profiles*. New York: John Wiley.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281-302.

Cunningham, W. R. (1981). Ability factor structure differences in adulthood and old age. *Multivariate Behavioral Research*, *16*, 3-22.

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

Drasgow, F. (1987). Study of measurement bias of two standardized psychological tests. *Journal of Applied Psychology, 72*, 19-29.
Drasgow, F., & Kanfer, R. (1985). Equivalence of psychological measurement in heterogeneous populations. *Journal of Applied Psychology, 70*, 662-680.
Dumka, L. E., Stoerzinger, H. D., Jackson, K. M., & Roosa, M. W. (1996). Examination of the cross-cultural and cross-language equivalence of the parenting self-agency measure. *Family Relations, 45*, 216-222.
Duncan, T. E., & Duncan, S. C. (1995). Modeling the processes of development via latent variable growth curve methodology. *Structural Equation Modeling, 2*, 178-213.
Duncan, T. E., Duncan, S. C., & Stoolmiller, M. (1994). Modeling developmental processes using latent growth structural equation modeling. *Applied Psychological Measurement, 18*, 343-354.
Durvasula, S., Andrews, J. C., Lysonski, S., & Netemeyer, R. G. (1993). Assessing the cross-national applicability of consumer behavior models: A model of attitude toward advertising in general. *Journal of Consumer Research, 19*, 626-636.
Embretson, S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*, 179-197.
Ferrando, P. J. (1996). Calibration of invariant item parameters in a continuous item response model using the extended LISREL measurement submodel. *Multivariate Behavioral Research, 31*, 419-439.
Finch, J. F., Okun, M. A., Barrera, M., Jr., Zautra, A. J., & Reich, J. W. (1989). Positive and negative social ties among older adults: Measurement models and the prediction of psychological distress and well-being. *American Journal of Community Psychology, 17*, 585-605.
Finch, J. F., & West, S. G. (1997). The investigation of personality structure: Statistical models. *Journal of Research in Personality, 31*, 439-485.
Golembiewski, R. T., Billingsley, K., & Yeager, S. (1976). Measuring change and persistence in human affairs: Types of change generated by OD designs. *Journal of Applied Behavioral Science, 12*, 133-157.
Gulliksen, H., & Wilks, S. S. (1950). Regression tests for several samples. *Psychometrika, 15*, 91-114.
Hattrup, K., Schmitt, N., & Landis, R. S. (1992). Equivalence of constructs measured by job-specific and commercially available aptitude tests. *Journal of Applied Psychology, 77*, 298-308.
Hertzog, C., & Schaie, K. W. (1986). Stability and change in adult intelligence: 1. Analysis of longitudinal covariance structures. *Psychology and Aging, 1*, 159-171.
Hinkin, T. R. (1995). A review of scale development practices in the study of organizations. *Journal of Management, 21*, 967-988.
Hinkin, T. R. (1998). A brief tutorial on the development of measures for use in survey questionnaires. *Organizational Research Methods, 1*, 104-121.
Hittner, J. B. (1995). Factorial validity and equivalency of the alcohol expectancy questionnaire tension-reduction subscale across gender and drinking frequency. *Journal of Clinical Psychology, 51*, 563-576.
Hofer, S. M., Horn, J. L., & Eber, H. W. (1997). A robust five-factor structure of the 16PF: Strong evidence from independent rotation and confirmatory factorial invariance procedures. *Personality & Individual Differences, 23*, 247-269.
Hong, E. (1995). A structural comparison between state and trait self-regulation models. *Applied Cognitive Psychology, 9*, 333-349.
Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research, 18*, 117-144.
Hu, L., & Bentler, P. M. (1993). Evaluating model fit. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 16-99). Newbury Park, CA: Sage.

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

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

James, L. R., Mulaik, S. A., & Brett, J. M. (1982). *Causal analysis: Models, assumptions, and data*. Beverly Hills, CA: Sage.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika, 34*, 183-202.

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

Jöreskog, K. G. (1974). Simultaneous analysis of longitudinal data from several cohorts. In W. M. Mason & S. E. Fienberg (Eds.), *Cohort analysis in social research* (pp. 323-341). New York: Springer Verlag.

Jöreskog, K., & Sörbom, D. (1996). *LISREL 8 user's guide*. Chicago: Scientific Software.

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

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.

Labouvie, E., & Ruetsch, C. (1995). Testing for equivalence of measurement scales: Simple structure and metric invariance reconsidered. *Multivariate Behavioral Research, 30*, 63-76.

Lance, C. E., Meade, A. W., & Williamson, G. M. (in press). We should measure change—and here's how. In G. M. Williamson & D. R. Shaffer (Eds.), *Physical illness and depression in older adults: Theory, research, and practice*. New York: Plenum.

Lance, C. E., Vandenberg, R. J., & Self, R. M. (1999, April). *Latent growth models of individual change: The case of newcomer adjustment*. Paper presented at the meeting of the Society for Industrial and Organizational Psychology, Atlanta, GA.

Li, F., Harmer, P., Acock, A., Vongjaturapat, N., & Boonverabut, S. (1997). Testing the cross-cultural validity of TEOSQ and its factor covariance and mean structures across gender. *International Journal of Sport Psychology, 28*, 271-286.

Li, F., Harmer, P., Chi, L., & Vongjaturapat, N. (1996). Cross-cultural validation of the task and ego orientation in sport questionnaire. *Journal of Sport and Exercise Psychology, 18*, 392-407.

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

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Manolis, C., Levin, A., & Dahlstrom, R. (1997). A Generation X scale: Creation and validation. *Educational and Psychological Measurement, 57*, 666-684.

Marcoulides, G. A. (1996). Estimating variance components in generalizabiilty theory: The covariance structure analysis approach. *Structural Equation Modeling, 3*, 290-299.

Marcoulides, G. A. (Ed.). (1998). *Applied generalizability theory models*. Mahwah, NJ: Lawrence Erlbaum.

Marcoulides, G. A., & Schumacker, R. E. (1996). *Advanced structural equation modeling: Issues and techniques*. Mahwah, NJ: Lawrence Erlbaum.

Marín, B. V., Gómez, C. A., Tschann, J. M., & Gregorich, S. E. (1997). Condom use in unmarried Latino men: A test of cultural constructs. *Health Psychology, 16*, 458-467.

Marsh, H. W. (1987). The factorial invariance of responses by males and females to a multidimensional self-concept instrument: Substantive and methodological issues. *Multivariate Behavioral Research, 22*, 457-480.

Marsh, H. W. (1993). The multidimensional structure of academic self-concept: Invariance over gender and age. *American Educational Research Journal, 30*, 841-860.

Marsh, H. W. (1994). Confirmatory factor analysis models of factorial invariance: A multifaceted approach. *Structural Equation Modeling, 1*, 5-34.

Marsh, H. W. (1995). The Δ^2 and χ^2I2 fit indices for structural equation models: A brief note of clarification. *Structural Equation Modeling, 2*, 246-254.

Marsh, H. W., Balla, J. R., & Hau, K. (1996). An evaluation of incremental fit indices: A clarification of mathematical and empirical properties. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling: Issues and techniques* (pp. 315-353). Mahwah, NJ: Lawrence Erlbaum.

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

Marsh, H., & Byrne, B. M. (1993). Confirmatory factor analysis of multitrait-multimethod self-concept data: Between-group and within-group invariance constraints. *Multivariate Behavioral Research, 28*, 313-349.

Marsh, H. W., & Hocevar, D. (1985). Application of confirmatory factor analysis to the study of self-concept: First- and higher order factor models and their invariance across groups. *Psychological Bulletin, 97*, 562-582.

Marsh, H. W., & Roche, L. A. (1996). Structure of artistic self-concepts for performing arts and non-performing arts students in a performing arts high school: "Setting the stage" with multigroup confirmatory factor analysis. *Journal of Educational Psychology, 88*, 461-477.

Maurer, T. J., Raju, N. S., & Collins, W. C. (1998). Peer and subordinate performance appraisal measurement equivalence. *Journal of Applied Psychology, 83*, 693-702.

McArdle, J. J. (1988). Dynamic but structural equation modeling of repeated measures data. In J. R. Nesselroade & R. B. Cattell (Eds.), *Handbook of multivariate experimental psychology* (2nd ed., pp. 561-614). New York: Plenum.

McArdle, J. J., & Aber, M. S. (1990). Patterns of change within latent variable structural equation models. In A. von Eye (Ed.), *Statistical methods in longitudinal research* (Vol. 1, pp. 151-224). Boston: Academic Press.

McArdle, J. J., & Anderson, E. (1990). Latent growth models for research on aging. In J. E. Birren & K. W. Schaie (Eds.), *Handbook of the psychology of aging* (3rd ed., pp. 21-44). San Diego, CA: Academic Press.

McDonald, R. P., & Marsh, H. W. (1990). Choosing a multivariate model: Noncentrality and goodness-of-fit. *Psychological Bulletin, 107*, 247-255.

Medsker, G. J., Williams, L. J., & Holahan, P. J. (1994). A review of current practices for evaluating causal models in organizational behavior and human resources management research. *Journal of Management, 20*, 439-464.

Meredith, W. (1964a). Notes on factorial invariance. *Psychometrika, 29*, 177-185.

Meredith, W. (1964b). Rotation to achieve factorial invariance. *Psychometrika, 29*, 187-207.

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

Millsap, R. E., & Everson, H. T. (1991). Confirmatory measurement model comparisons using latent means. *Multivariate Behavioral Research, 26*, 479-497.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17*, 297-334.

Millsap, R. E., & Hartog, S. B. (1988). Alpha, beta, and gamma change in evaluation research: A structural equation approach. *Journal of Applied Psychology, 73*, 574-584.

Mulaik, S. A. (1972). *The foundations of factor analysis*. New York: McGraw-Hill.

Mulaik, S. A. (1975). Confirmatory factor analysis. In D. J. Amick & H. J. Walberg (Eds.), *Introductory multivariate analysis* (pp. 170-207). Berkeley, CA: McCutchan.

Mulaik, S. A., James, L. R., Van Alstine, J., Bennett, N., Lind, S., & Stilwell, C. D. (1989). Evaluation of goodness-of-fit indices for structural equation models. *Psychological Bulletin, 105*, 430-445.

Nesselroade, J. R., & Thompson, W. W. (1995). Selection and related threats to group comparisons: An example comparing factorial structures of higher and lower ability groups of adults twins. *Psychological Bulletin*, *117*, 271-284.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

O'Reilly, C., III, & Chatman, J. (1986). Organizational commitment and psychological attachment: The effects of compliance, identification, and internalization of prosocial behavior. *Journal of Applied Psychology*, *71*, 492-499.

Palich, L. E., Hom, P. W., & Griffeth, R. W. (1995). Managing in the international context: Testing cultural generality of sources of commitment to multinational enterprises. *Journal of Management*, *21*, 671-690.

Parker, C. P., Baltes, B. B., & Christiansen, N. D. (1997). Support for affirmative action, justice perceptions, and work attitudes: A study of gender and racial-ethnic group differences. *Journal of Applied Psychology*, *82*, 376-389.

Pentz, M. A., & Chou, C. (1994). Measurement invariance in longitudinal clinical research assuming change from development and intervention. *Journal of Consulting and Clinical Psychology*, *62*, 450-462.

Pike, G. (1991). Using structural equation models with latent variables to study student growth and development. *Research in Higher Education*, *32*, 499-524.

Pike, G. (1996). Limitations of using students' self-reports of academic development as proxies for traditional achievement measures. *Research in Higher Education*, *37*, 89-114.

Pitts, S., West, S., & Tein, J. (1996). Longitudinal measurement models in evaluation research: Examining stability and change. *Evaluation and Program Planning*, *19*, 333-350.

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

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

Rock, D. A., Werts, C. E., & Flaugher, R. L. (1978). The use of analysis of covariance structures for comparing the psychometric properties of multiple variables across populations. *Multivariate Behavioral Research*, *13*, 403-418.

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.

Schaie, K. W., & Hertzog, C. (1985). Measurement in the psychology of adulthood and aging. In J. E. Birren & K. W. Schaie (Eds.), *Handbook of the psychology of aging* (2nd ed., pp. 61-92). New York: Van Nostrand Reinhold.

Schaie, K., Maitland, S., Willis, S., & Intrieri, R. (1998). Longitudinal invariance of adult psychometric ability factor structures across 7 years. *Psychology and Aging*, *3*, 8-20.

Schaubroeck, J., & Green, S. G. (1989). Confirmatory factor analytic procedures for assessing change during organizational entry. *Journal of Applied Psychology, 74*, 892-900.

Schmidt, F. L., & Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods*, *1*, 199-223.

Schmit, M. J., Ryan, A. M., Stierwalt, S. L., & Powell, A. B. (1995). Frame-of-reference effects on personality scale scores and criterion-related validity. *Journal of Applied Psychology*, *80*, 607-620.

Schmitt, N. (1982). The use of analysis of covariance structures to assess beta and gamma change. *Multivariate Behavioral Research*, *17*, 343-358.

Schmitt, N., Pulakos, E. D., & Lieblein, A. (1984). Comparison of three techniques to assess group-level beta and gamma change. *Applied Psychological Measurement*, *8*, 249-260.

Schulenberg, J. E., Shimizu, K., Vondracek, F. W., & Hostetler, M. (1988). Factorial invariance of career indecision dimensions across junior high and high school males and females. *Journal of Vocational Behavior, 33*, 63-81.

Singh, J. (1995). Measurement issues in cross-national research. *Journal of International Business Studies*, *26*, 597-619.

Smith, C. S., Tisak, J., Bauman, T., & Green, E. (1991). Psychometric equivalence of a translated circadian rhythm questionnaire: Implications for between-and within-population assessments. *Journal of Applied Psychology, 76*, 628-636.

Soeken, K. L., & Prescott, P. A. (1991). Patient intensity for nursing index: The measurement model. *Research in Nursing & Health*, *14*, 297-304.

Stacy, A. W., MacKinnon, D. P., & Pentz, M. A. (1993). Generality and specificity in health behavior: Application to warning-label and social influence expectancies. *Journal of Applied Psychology, 78*, 611-627.

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

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

Steiger, J. H., Shapiro, A., & Browne, M. W. (1985). On the multivariate asymptotic distribution of sequential chi-square statistics. *Psychometrika*, *50*, 253-264.

Tanaka, J. S. (1993). Multifaceted conceptions of fit in structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 10-39). Newbury Park, CA: Sage.

Tansy, M., & Miller, J. A. (1997). The invariance of the self-concept construct across White and Hispanic student populations. *Journal of Psychoeducational Assessment*, *15*, 4-14.

Taris, T. W., Bok, I. A., & Meijer, Z. Y. (1998). Assessing stability and change of psychometric properties of multi-item concepts across different situations: A general approach. *Journal of Psychology, 132*, 301-316.

te Nijenhuis, J., & van der Flier, H. (1997). Comparability of GATB scores for immigrants and majority group member: Some Dutch findings. *Journal of Applied Psychology*, *82*, 675-687.

Terborg, J. R., Howard, G. S., & Maxwell, S. E. (1980). Evaluating planned organizational change: A method for assessing alpha, beta, and gamma change. *Academy of Management Review, 5*, 109-121.

Thurstone, L. L. (1947). *Multiple-factor analysis*. Chicago: University of Chicago Press.

Tucker, L. R., & Lewis, C. (1973). The reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, *38*, 1-10.

Van den Bergh, B.R.H., & Van Ranst, N. (1998). Self-concept in children: Equivalence of measurement and structure across gender and grade of Harter's self-perception profile for children. *Journal of Personality Assessment*, *70*, 564-582.

Van de Vijver, F.J.R., & Harsveld, M. (1994). The incomplete equivalence of the paper-and-pencil and computerized versions of the general aptitude test battery. *Journal of Applied Psychology, 79*, 852-859.

Van Dyne, L., & LePine, J. A. (1998). Helping and voice extra-role behaviors: Evidence of construct and predictive validity. *Academy of Management Journal*, *41*, 108-119.

Van Ranst, N., & Marcoen, A. (1997). Meaning in life of young and elderly adults: An examination of the factorial validity and invariance of the life regard index. *Personality and Individual Differences*, *22*, 877-884.

Vandenberg, R. J., & Scarpello, V. (1990). The matching model: An examination of the processes underlying realistic job previews. *Journal of Applied Psychology*, *75*, 60-67.

Vandenberg, R. J., & Self, R. M. (1993). Assessing newcomers' changing commitment to the organization during the first 6 months of work. *Journal of Applied Psychology*, *78*, 557-568.

Vandenberg, R. J., Self, R. M., & Seo, J. H. (1994). A critical examination of the internalization, identification, and compliance commitment measures. *Journal of Management*, *20*, 123-140.

Veerman, J. W., ten Brink, L. T., Straathof, M.A.E., & Treffers, P.D.A. (1996). Measuring children's self-concept with a Dutch version of the "self-perception profile for children": Factorial validity and invariance across a nonclinic and a clinic group. *Journal of Personality Assessment*, *67*, 142-154.

Willett, J. B., & Sayer, A. G. (1994). Using covariance structure analysis to detect correlates and predictors of individual change over time. *Psychological Bulletin*, *116*, 363-381.

Windle, M., Iwawaki, S., & Lerner, R. M. (1988). Cross-cultural comparability of temperament among Japanese and American preschool children. *International Journal of Psychology, 23*, 547-567.

Robert J. Vandenberg received his Ph.D. in social psychology from the University of Georgia and is an associate professor in the Department of Management of the Terry College of Business at the University of Georgia. He is currently division chair elect of the Research Methods Division of the Academy of Management and is on the editorial boards of the Journal of Applied Psychology, Organizational Behavior and Human Decision Processes, *and* Organizational Research Methods. *His research interests include research methods, high involvement work processes, and employee work adjustment processes.*

Charles E. Lance received his Ph.D. in industrial and organizational psychology from the Georgia Institute of Technology and is professor and chair of the applied psychology program at the University of Georgia. He is a former president of the Atlanta Society of Applied Psychology and is on the editorial boards of Groups and Organization Management, Human Resource Management Review, Organizational Research Methods, *and* Personnel Psychology. *His research interests include research methods and statistics, performance measurement, and personnel psychology.*