Extreme response styles (ERS) and acquiescence response styles (ARS) may constitute important sources of cross-cultural differences on survey-type instruments. Differences in ERS and ARS, if undetected, may give rise to spurious results that do not reflect genuine differences in attitudes or perceptions. Multiple-group confirmatory factor analysis is recommended as the most effective method of testing for ERS and ARS and determining whether cultural groups can be meaningfully compared on the basis of factor (latent) means. A detailed numerical example is provided.

ASSESSING EXTREME AND ACQUIESCENCE RESPONSE SETS IN CROSS-CULTURAL RESEARCH USING STRUCTURAL EQUATIONS MODELING

GORDON W. CHEUNG
The Chinese University of Hong Kong

ROGER B. RENSVOLD
City University of Hong Kong

Cross-cultural research is becoming increasingly important owing to such factors as the heterogeneity of the international workforce, the expansion of global markets, and the increasing influence of multinational corporations (Triandis, 1994). One indicator of this increased interest is the most recent edition of the *Handbook of Industrial and Organisational Psychology*, which dedicates an entire volume to the subject. Cross-cultural topics are also appearing more frequently in psychology and management journals.

Cross-cultural studies usually hypothesize culturally based differences in individual perceptions and attitudes. Unfortunately, however, observed differences may be due to one or more measurement artifacts unrelated to the constructs of interest (Adler, 1984; Berry, Poortinga, Segall, & Dasen, 1992; Irvine & Carroll, 1980; Mullen, 1995; Poortinga, 1989; Singh, 1995; Triandis, 1994; van de Vijver & Leung, 1997; van de Vijver & Poortinga, 1982). It is important that researchers have access to appropriate statistical tools for identifying and interpreting such artifacts, particularly in cross-cultural work.

AUTHORS' NOTE: Both authors contributed equally to the preparation of this manuscript. The order of authorship is alphabetical. An earlier version of this article, *Inside the Black Box of Cross-Cultural Comparisons*, coauthored by Gordon W. Cheung and Christine M. Chan, was presented at the International Management Division, 1997 Annual Meeting of the Academy of Management in Boston.

JOURNAL OF CROSS-CULTURAL PSYCHOLOGY, Vol. 31 No. 2, March 2000 187-212

Recent discussions of cross-cultural measurement artifacts have addressed various forms of invariance and noninvariance in the measurement model. Conceptual equivalence, operationalized as factor form invariance (e.g., Buss & Royce, 1975; Irvine, 1969; Suzuki & Rancer, 1994), exists when members of both cultures associate the same measures (e.g., survey items) with the same underlying factors. Factorial invariance (e.g., Drasgow, 1984; Drasgow & Kanfer, 1985; Janssens, Brett, & Smith, 1995; Meredith, 1993; Reise, Widaman, & Pugh, 1993; Smith, Tisak, Bauman, & Green, 1991) exists when members of both cultures ascribe approximately the same weight to indicators, as manifested by equal (strictly speaking, not significantly different) factor loading parameters. Others have extended the discussion to include measurement error equivalence (Mullen, 1995; Singh, 1995) and calibration equivalence (Riordan & Vandenberg, 1994). Little (1997) provides a recent discussion of cross-cultural measurement artifacts, in the context of the analysis of covariance.

The various forms of measurement noninvariance are important to cross-cultural research, yet the naïve approach (which is becoming less common) ignores them completely. The naïve approach involves taking a scale that has been validated in Culture A, administering it in Culture B, and then uncritically comparing the scale scores using a *t* test or some similar procedures. The observed difference, if there is one, may be due to a cultural difference; however, it may also be due to one or more invariance failures in the measurement model.

This article reviews procedures for testing cross-cultural measurement models for noninvariance using structural equations modeling (SEM) methodology. These procedures are not new. The article's contribution is to demonstrate that certain types of noninvariance may be interpreted as manifestations of two well-known response set biases, namely, extreme response style (ERS) and acquiescence response style (ARS). Demonstrating that a measurement model is free of ERS and ARS eliminates alternative explanations for observed cross-cultural differences. Conversely, demonstrating the existence of ERS or ARS either adds caveats to the interpretation of a cross-cultural measurement, or (in some instances) suggests how the measurement model can be "fixed." The procedures themselves may prove enlightening to researchers who are thoroughly familiar with ERS and ARS, but not with SEM. In addition, the SEM approach is an attractive alternative to other tests for ERS and ARS, some of which require large samples and/or large numbers of items (e.g., Chun, Campbell, & Yoo, 1974; Cunningham, Cunningham, & Green, 1977).

There is, of course, no suggestion that all invariance failures are attributable to either ERS or ARS. They may, for example, be due to undetected errors in translation, coding blunders, or inappropriate sampling procedures. A statistical finding suggesting the existence of a between-group difference in either ERS or ARS needs to be supported by the investigator's understanding of the research context.

ERS

ERS is the tendency to use the extreme categories of rating scales. If high-ERS participants are given a survey using a 7-point Likert-type scale, their responses will tend to be either 1 (*strongly agree*) or 7 (*strongly disagree*). If low-ERS participants are given the same survey, their responses will tend to cluster around 4 (*neither agree nor disagree*). Equivalence of ERS is sometimes referred to as calibration equivalence, meaning that both groups employ the same scale of measurement (Berry et al., 1992; Mullen, 1995). Several studies have documented cross-cultural differences in ERS (Greenleaf, 1992; Hui & Triandis, 1985; Schaninger & Buss, 1986; Triandis, 1994). Lee and Green (1991), for example, discovered that Koreans tend to avoid extremes and prefer the midpoints of scales. Differences may be due to culturally based response norms (Guptara, Murray, Razak, & Sheehan, 1990; Hui & Triandis, 1989; Marín, Gamba, & Marín, 1992; Zax & Takahashi, 1967). Members of low-ERS cultures may desire to appear modest and nonjudgmental, whereas members of high-ERS cultures may wish to demonstrate sincerity and conviction. Alternatively, cultural differences may affect how items are interpreted (Riordan & Vandenberg, 1994). One culture may have no particular opinion concerning the content of a survey item, whereas another culture may have strong opinions that are, in addition, highly polarized. This would lead to a cross-cultural difference in ERS with respect to that item.

Some researchers (Greenleaf, 1992; Hui & Triandis, 1985) identify ERS with scale standard deviations. Strictly speaking, ERS is not identical with the standard deviation (Greenleaf, 1992) although it is highly correlated with it.

ERS differences have many adverse effects on cross-cultural and international comparisons (Chun, Campbell, & Yoo, 1974). Because ERS affects numerical scores, comparisons of means become uninterpretable. At a more fundamental level, ERS differences produce noninvariant factor loadings and intercepts (as discussed below), leading to the conclusion that the numbers on the response scale mean different things to members of different groups.

Between-group differences in ERS may be either nonuniform or uniform. In a nonuniform ERS difference, only a subset of items are affected. In uniform ERS, all items are affected. The two cases are discussed separately below.

ARS

An ARS difference, sometimes referred to as scalar nonequivalence (Mullen, 1995), occurs when one group systematically gives higher or lower responses than another group, resulting in a scale displacement. Several studies have documented cross-cultural ARS differences (Cunningham, Cunningham, & Green, 1977; England & Harpaz, 1983; Morris & Pavett, 1992). For example, Riordan and Vandenberg (1994) found that a response of 3 on a 5-point Likert-type scale means "no opinion" to American respondents but "mild agreement" to Korean respondents. As a result of this scale displacement, Korean "3"s were equivalent to American "4"s and Korean "4"s were equivalent to American "5"s.

Cross-cultural differences in ARS can be explained in terms of social desirability, a belief that a higher score is a better score, or by a preoccupation with individual defects and deficiencies (Guilford, 1954; Hui & Triandis, 1985; Moorman & Podsakoff, 1992; Peterson & Wilson, 1992). On the individual level, some respondents display extreme ARS by agreeing (or disagreeing) with almost any statement (Guilford, 1954; Peterson & Wilson, 1992; Triandis, 1994). Like ERS, a cross-cultural ARS difference can be either nonuniform (affecting some responses) or uniform (affecting all responses). The two cases are considered separately below.

THE MEASUREMENT MODEL: EQUIVALENCE REQUIREMENTS[1]

This section reviews the elements of the measurement model and describes the requirements that must be satisfied before the model can be used to compare groups on the basis of latent means (Bollen, 1989).

An example of the general measurement model is shown in Figure 1. Variances in the manifest variables (e.g., item responses) X_i (i = 1 through 6) are the result of variance in the latent variables (constructs) $\xi_{(j\,=\,1,\,2)}$, plus error terms δ_i. Construct variances are ϕ_{11} and ϕ_{22}, and the covariance is ϕ_{21}. The factor loading parameters λ_{ij} represent the strength of the relationships between each construct and its associated items (Bollen, 1989; Jöreskog & Sörbom, 1993). It is convenient to think of λ_{ij} as the slopes of regression lines, that is, as the weights obtained by regressing the item responses on the constructs.

A two-group case of the Figure 1 model is shown in the left panel of Figure 2, with group membership indicated by parenthetical superscripts. The first requirement for comparing groups is *form invariance*. When the model is estimated for each group, the same items must be associated with each

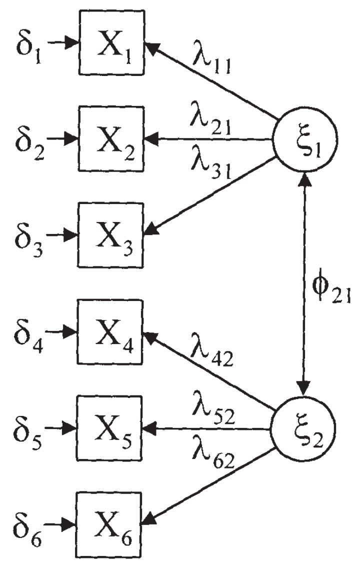


Figure 1: General Measurement Model (example)

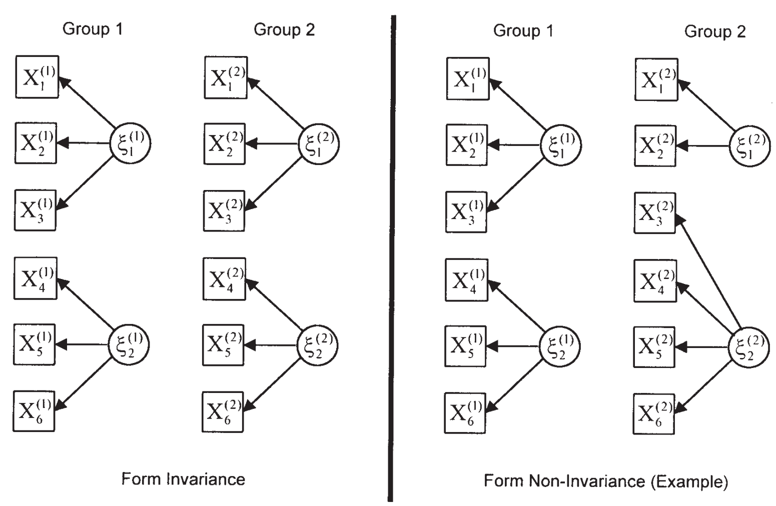


Figure 2: Form Invariance and Noninvariance

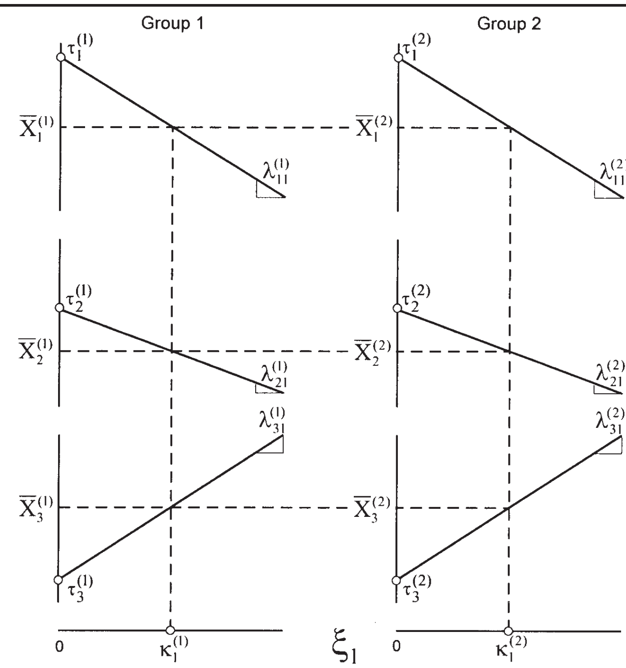
construct. The left panel of Figure 2 shows form invariance and the right panel shows one of many possible instances of form noninvariance.

Statistical tests for this and other types of invariance are described below; however, it should be noted at this point that one type of form noninvariance, known as construct bias (van de Vijver & Poortinga, 1997), cannot be detected statistically. Some constructs have wider scope in one culture than in another; filial piety, for example, is a more highly elaborated construct in China than it is in the West (Hsieh, 1967). As a result, measuring this construct with adequate validity may require more items in one culture than it does in the other. Therefore, a particular set of items may be conceptually adequate for assessing a construct in one culture, inadequate in a second culture, and yet display form invariance when compared using data from both cultures. To avoid this type of bias, a researcher should construct scales using one or more of the culturally based approaches described by van de Vijver and Leung (1997).

Groups are frequently compared on the basis of scale scores. A scale score can be calculated in many ways, but usually as the sum, mean, or weighted mean of item responses. In some instances, however, the preferred datum for group comparisons is the latent mean. The latent mean is estimated as part of a structural equations model, which includes the error terms, and this is its principle advantage. Estimating error terms improves the estimate of the mean (decreases its standard error) by partialing out variance attributable to measurement error. The latent mean of a construct is symbolized by a lowercase kappa, with a subscript indicating the construct, and (in our nomenclature) a parenthetical superscript indicating group. For example, $\kappa_1^{(2)}$ represents the latent mean of ξ_1, estimated using Group 2 data.

Consider construct ξ_1 in the left panel of Figure 2. Variance in ξ_1 produces variance in responses to items X_1, X_2, and X_3. Suppose that a researcher has collected Group 1 and Group 2 responses. If the between-group differences $\Delta\overline{X}_i$ ($\Delta\overline{X}_i = \overline{X}_i^{(1)} - \overline{X}_i^{(2)}, \mathrm{i} = 1, 2, 3$) are not significantly different, then it appears that the latent means $\kappa_1^{(1)}$ and $\kappa_1^{(2)}$ of the underlying construct are not significantly different. This, however, is not necessarily true.

The invariance requirements for equality of latent means are shown in Figure 3. In addition to form invariance (see Figure 2), the values of the λ_{ij} must be invariant across groups ($\lambda_{11}^{(1)} = \lambda_{11}^{(2)}$, etc.).This condition is referred to as factorial invariance (Bollen, 1989; Jöreskog & Sörbom, 1993; Meredith, 1993). In addition the intercept of each item, i.e., the value of the item corresponding to $\xi_1 = 0$, must also be invariant ($\tau_1^{(1)} = \tau_1^{(2)}$, etc.). The consequences of noninvariance are shown in Figure 4. In the left-hand panel, factorial noninvariance of item X_3 ($\lambda_{31}^{(1)} = \lambda_{31}^{(2)}$) implies that the same level of $\overline{X}_3$ corresponds to two different values of κ_1. In the right-hand panel, intercept



$$\begin{bmatrix} \overline{X}_1^{(1)} = \overline{X}_1^{(2)} \\ \overline{X}_2^{(1)} = \overline{X}_2^{(2)} \\ \overline{X}_3^{(1)} = \overline{X}_3^{(2)} \end{bmatrix} \text{ implies } \kappa_1^{(1)} = \kappa_1^{(2)} \text{ if and only if } \begin{bmatrix} \lambda_{11}^{(1)} = \lambda_{11}^{(2)} \\ \lambda_{21}^{(1)} = \lambda_{21}^{(2)} \\ \lambda_{31}^{(1)} = \lambda_{31}^{(2)} \end{bmatrix} \& \begin{bmatrix} \tau_1^{(1)} = \tau_1^{(2)} \\ \tau_2^{(1)} = \tau_2^{(2)} \\ \tau_3^{(1)} = \tau_3^{(2)} \end{bmatrix}$$

-- in addition to form invariance (Figure 2).

Figure 3: Invariance Requirements for Comparison of Latent Means

noninvariance of item X_2 ($\tau_2^{(1)} \neq \tau_2^{(2)}$) implies that the same value of $\overline{X}_2$ corresponds to two different values for κ_1

Alternatively, one can speak of the difference rather than the equality of latent means. A typical research objective is to test the existence of a predicted between-group difference. If survey items do not display form and factorial and intercept invariance, then it cannot be determined whether the observed difference is attributable to the hypothesized difference in the construct or rather is an artifact of noninvariance.

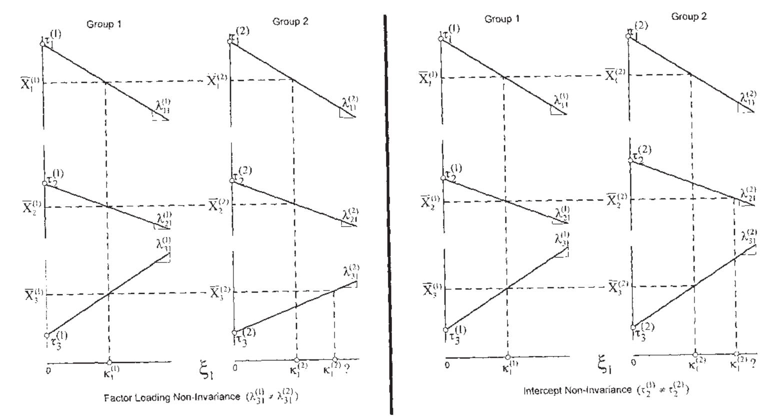


Figure 4: Consequences of Noninvariance

ERS AND ARS IN THE MEASUREMENT MODEL

We present a contrived example to show the effects of ERS and ARS on factor loadings and intercepts. Consider 4 groups of 40 participants each (see Table 1). It has been determined, using an (imaginary) error-free instrument, that the four groups are identical with respect to a latent variable (e.g., job satisfaction). Ten members of each group are low with respect to the construct, twenty are midrange, and ten are high. The four groups differ, however, with respect to ERS and ARS.

The four groups respond to a single positively worded item (e.g., "I enjoy going to work in the morning") using a 5-point Likert-type scale. The members of Group 1 are high in ERS and tend to overstate their attitudes, whereas the members of Group 2 are low in ERS and tend to understate. The difference between the groups appears in the distribution of extreme responses. For example, within the 10 members of each group who are lowest in job satisfaction, 6 members of Group 1 versus no members of Group 2 select response 1 (*strongly disagree*). The slopes of the regression lines (factor loadings) are 1.5 for Group 1 versus 0.5 for Group 2 (see Figure 5). The intercepts are also different: 0.0 for Group 1 versus 2.0 for Group 2.

Groups 3 and 4 in Table 1 are the same as Group 2 with respect to ERS but exhibit contrasting levels of ARS. The members of Group 3 are high in ARS, or "yeasayers." They are consistently biased in favor of a positive response; in fact, their responses are exactly one point higher than those of Group 2.

TABLE 1
Levels of ERS and ARS
(responses to hypothetical item X_1; groups of 40)

		Levels of Latent Variable		
		1 (Low)	*2 (Mid)*	*3 (High)*
		Number of participants at each level		
	Response (1)	*10*	*20*	*10*
	1	6		
Group 1: High ERS "Overstatement"	2	3	5	
	3	1	10	1
	4		5	3
	5			6
	1			
Group 2: Low ERS "Understatement"	2	5	2	
	3	5	16	5
	4		2	5
	5			
	1			
Group 3: High ARS "Yeasayers"	2			
	3	5	2	
	4	5	16	5
	5		2	5
	1	5	2	
Group 4: Low ARS "Naysayers"	2	5	16	5
	3		2	5
	4			
	5			

(1) Responses: 1 = *strongly disagree*, 2 = *disagree*, 3 = *neither agree nor disagree*, 4 = *agree*, 5 = *strongly agree*.

Group 4, consisting of "naysayers," are biased to the same degree but in the opposite direction. The factor loadings (see Figure 5) of Groups 3 and 4 are both 0.5, the same as Group 2. The intercepts, however, are different: 3.0 for Group 3 versus 1.0 for Group 4.

Extreme ARS due to the tone of an item may also produce a failure of form invariance; that is, members of different cultures may associate the item with different constructs. It is possible for participants to respond strictly to tone, disregarding the underlying meaning of an item. Sensitivity to tone, however, may be an artifact of culture. For example, the wording of an item may seem reasonable to an American respondent but strident and hostile to a Korean respondent, and therefore trigger an extreme level of ARS. The American participants would respond to the content of the item, whereas the Koreans

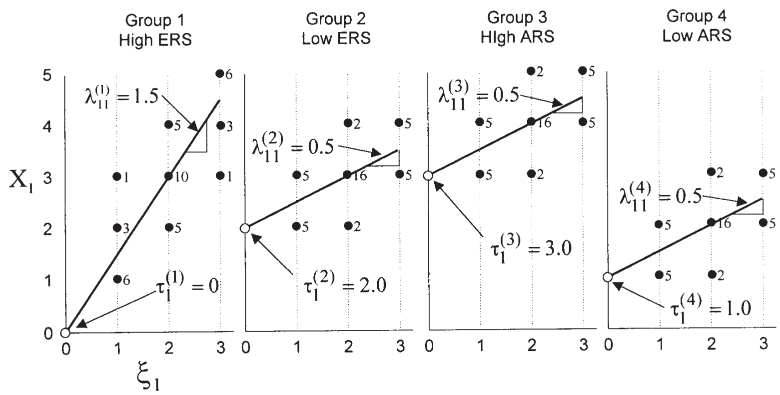


Figure 5: Effects of Extreme Response Style (ERS) and Acquiescence Response Style (ARS) on Factor Loadings and Intercepts (data from Table 1)

would respond principally to its tone. In the Korean sample, therefore, participant responses would tend to be unrelated to the underlying construct, and the Korean factor loading would therefore be lower than the American loading. This may cause failure of either form invariance or factorial invariance. In the first instance, the Korean loading is not significantly different from zero, whereas the American loading is nonzero; therefore, it is not true that both groups associate the same set of items with the same construct. In the second instance, both loadings are nonzero but are significantly different in magnitude.

In summary, ERS and ARS are associated with factorial noninvariance (leading in extreme cases to form noninvariance) and intercept noninvariance. ERS affects both factor loadings and intercepts, whereas ARS affects intercepts. Testing for factorial invariance—that is, examining factor loadings across groups for statistically significant differences—can be used to determine whether groups display different levels of ERS with respect to particular items. It cannot, however, detect differences in ARS. To accomplish this task, the researcher must explicitly test for equality of intercepts. We now consider the details of these tests.

TESTS FOR INVARIANCE USING SEM

Table 2 presents the sequence of tests for invariance in summary form (Bollen, 1989; Jöreskog & Sörbom, 1989; Mullen, 1995; Singh, 1995). The

TABLE 2
Sequence of Tests for Invariance of Latent Means (perceived aspects of job quality, Figure 6)

H(#)	*Symbol*	*Constraint*
H(1)	H_{form}	Form invariance (Figure 2). Adequate fit when models for both groups are estimated simultaneously using the same factor form.
H(2)	H_{Λ_x}	$\lambda_{ij}^{(1)} = \lambda_{ij}^{(2)}$ for all λ_{ij} in the model.
H(3)	$H_{\Lambda_x \nu_x}$	H(2), plus $\tau_i^{(1)} = \tau_i^{(2)}$ for all τ_i in the model.
H(3.1)	$H_{\Lambda_x \nu_1}$	H(2), plus $\tau_i^{(1)} = \tau_i^{(2)}$ (Intercepts of items X_i associated with construct ξ_1; i = 1, 2, 3, 4)
H(3.2)	$H_{\Lambda_x \nu_2}$	H(2), plus $\tau_i^{(1)} = \tau_i^{(2)}$) (Intercepts of items X_i associated with construct ξ_2; i = 5, 6, 7, 8)
H(3.3)	$H_{\Lambda_x \nu_3}$	H(2), plus $\tau_i^{(1)} = \tau_i^{(2)}$ (Intercepts of items X_i associated with construct ξ_3; i = 9, 10, 11, 12)
H(4.1)	$H_{\Lambda_x \nu_1 \kappa_1}$	H(3.1), plus $\kappa_1^{(1)} = \kappa_1^{(2)}$
H(4.2)	$H_{\Lambda_x \nu_2 \kappa_2}$	H(3.2), plus $\kappa_2^{(1)} = \kappa_2^{(2)}$
H(4.3)	$H_{\Lambda_x \nu_3 \kappa_3}$	H(3.3), plus $\kappa_3^{(1)} = \kappa_3^{(2)}$

NOTE: $H_{(\#)}$ = the set of hypotheses characterizing the constrained model; used also to represent the test itself.

details in the table (ranges of indices, etc.) reflect the characteristics of the model shown in Figure 6. This model also provides a numerical example later in the article.

The first test, for form invariance, hypothesizes an adequate fit when models for both groups are estimated simultaneously using the same factor structure, H(1). Failure to obtain an adequate fit suggests that group-specific patterns of factor loadings (as shown, for example, in the right panel of Figure 2) fit the data better than one overall pattern (see the left panel of Figure 2). If an adequate fit is not obtained at this point, then the attempt to compare latent means must be abandoned.

All subsequent tests involve comparing a constrained model with an unconstrained model. Each constrained model is estimated, as it is subject to the requirement that some set of parameters, such as the factor loadings, *must be equal* for both (or all) groups. The fit statistics of the constrained model are compared with those of the corresponding unconstrained model, which does not include the equality requirement. If differences in fit statistics indicate that the constrained model fits the data significantly *less well* than the unconstrained model, then the constrained parameters are noninvariant. This interpretation derives from the fact that the model fits the data better if the parameters are *not* forced to be the same for all groups. (The question of whether a difference in fit statistics is large enough to indicate noninvariance will be discussed in connection with the numerical example below.)

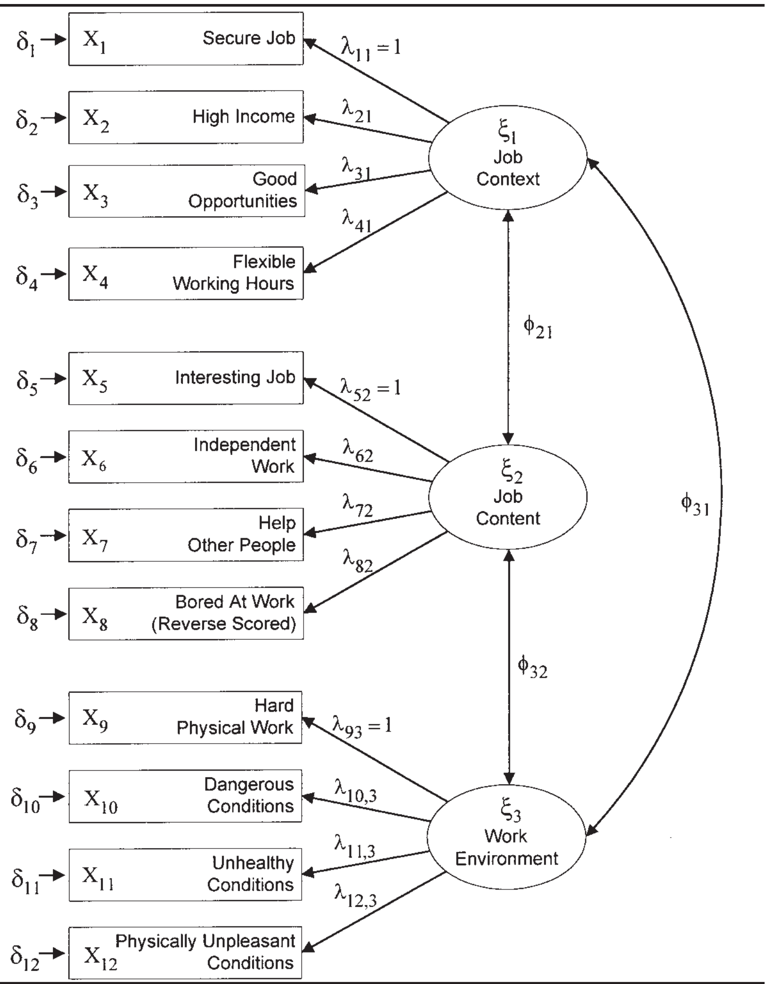


Figure 6: Perceived Aspects of Job Quality (model)

TESTING FOR ERS

H(2), the test for factorial invariance, hypothesizes that factor loadings are invariant. (Strictly speaking, H(2) is not the test itself, but rather the set of hypotheses characterizing the constrained model. In the interest of brevity, we will use the terminology interchangeably.) Factorial invariance exists if there is no significant difference in fit between H(2) and H(1). If factorial invariance is rejected, noninvariant items must be identified using an iterative procedure that is treated in detail elsewhere (Byrne, Shavelson, & Muthén, 1989; Cheung & Rensvold, (1999); Rensvold & Cheung, 1998). As noted above, factorial noninvariance suggests but does not prove the existence of a between-group difference in ERS. Items displaying factorial noninvariance should be examined to determine whether the ERS interpretation is reasonable.

Once identified, noninvariant items may be dropped from the model, if doing so does no substantive damage to construct validity or theory (Poortinga, 1989). Alternatively, if the items constitute only a small part of the scale and are believed to have no substantive effect, they may be retained under the doctrine of partial factorial invariance (Byrne, 1993; Marsh & Hocevar, 1985; Reise, Widaman, & Pugh, 1993; Riordan & Vandenberg, 1994). In any event, items having noninvariant factor loadings must either be dealt with in some way, or the attempt to compare latent means must be abandoned at this point.

In the best possible situation, factorial invariance holds for all the items associated with a particular construct. In the next best situation, it holds for most items, and the relatively few noninvariant items can either be eliminated (Poortinga, 1989) or retained under the doctrine of partial factorial invariance (e.g., Byrne, 1993). This can be referred to as a *nonuniform* ERS difference, since it exists with respect to only a few items and not to all of the items associated with the construct. As more items are found to be noninvariant, the situation becomes more intractable. A finding of factorial noninvariance for many or most items may indicate several things: the construct may be poorly operationalized; the data collection may have been flawed, such as through the use of inaccurately translated items; or there may be very strong cross-cultural differences in how the construct is conceptualized.

Unfortunately, a finding of factorial invariance does not rule out the possibility of a *uniform* cross-cultural difference in ERS. This condition is found when the same bias exists with respect to all items that serve as indicators of a construct. Referring to the left panel of Figure 2, suppose that Group 2 participants tend to give more extreme responses to items X_1, X_2, and X_3 than Group 1 participants, and that this tendency is the same for all three items. Then the factor loading parameters will not be significantly different across groups ($\lambda_{i1}^{(1)} = \lambda_{i1}^{(2)}$, i = 1, 2, 3). The variance of the latent variable ξ_1 will, however, be greater in Group 2 than in Group 1 ($\phi_{11}^{(2)} > \phi_{11}^{(1)}$). It should be noted, however, that inequality of variances is the expected state of affairs, because the two samples are drawn from different populations. Therefore, although $\phi_{11}^{(1)} = \phi_{11}^{(2)}$ rules out a uniform ERS difference, the converse $\phi_{11}^{(2)} \neq \phi_{11}^{(1)}$ does not necessarily demonstrate its presence.

Figure 5 shows that between-group ERS equivalence is a prerequisite for between-group ARS. Group 3 and Group 4 (see Figure 5) would not differ in ARS if the intercepts $\tau_1^{(3)}$ and $\tau_1^{(4)}$ were equal, yet this could only be attributed to coincidence if the slopes $\lambda_{11}^{(3)}$ and $\lambda_{11}^{(4)}$ were *not* equal (i.e., if equivalent ERS did not exist). If between-group differences in ERS can either be ruled out or eliminated by judiciously removing items, then it is appropriate to test for between-group differences in ARS.

TESTING FOR ARS

H(3) is the test for overall intercept invariance (i.e., equivalent ARS). Construct-level constrained models, H(3.1) through H(3.3), see Table 2, are also judged at this point to produce estimates of the latent means. Overall intercept invariance is tested by comparing the fit indices of H(3) and H(2). If overall invariance is rejected, then construct-level tests are conducted, for example H(3.1) versus H(2). If a construct displays intercept noninvariance, then the attempt to compare the latent means of that construct across groups must be abandoned.[2]

As was the case with ERS, the outcome of these tests may suggest but cannot prove the existence of a between-group difference in ARS. The items serving as indicators of the construct in question should be examined to determine whether the ARS explanation is reasonable in the context of the study.

The question of nonuniform versus uniform ERS arose in the previous section, and a similar question also arises with respect to ARS. If a between-group difference in ARS is uniform with respect to all the indicators of a construct, will a construct-level test identify the ARS difference? The answer is yes, unless all the factor loadings are identically equal to 1; a situation which is unlikely to be observed in practice. (A proof is available from the authors.)

The final series of tests, H(4.1) through (4.3), see Table 2, determine whether the latent means estimated in conjunction with H(3.1) through H(3.3) are equivalent. Nonequivalence is indicated by a significant difference between the fit statistics of the model with construct-level intercept constraints, for example H(3.1) and the same model with mean constraints added, H(4.1). A finding of nonequivalence at this point indicates that the between-group difference in the latent mean is *due to a substantive between-group difference* and not to a difference in either ERS or ARS.

AN EXAMPLE

We analyze a subset of data from the 1989 "work orientation" module of the International Social Survey Program (ISSP, 1989). The study included data from eleven countries and covered three main topics: (a) general attitudes towards work and leisure, (b) work organization, and (c) work content. Responses relating to the constructs' "quality of job context," "quality of job content," and "quality of the work environment" from participants in the United States (N=823) and Italy (N=548) are used in this example. Relationships between the constructs and indicators (item responses) are shown in Figure 6. Four indicators are associated with each construct; X_1 through X_4 with ξ_1 (job context), X_5 through X_8 with ξ_2 (job context), and X_9 through X_{12} with ξ_3 (work environment). A parenthetical superscript of (1) indicates a parameter associated with the U.S. sample, whereas a superscript of (2) refers to the Italian sample. Data were input in the form of Pearson correlation matrices, matrices of observed means, and matrices of standard deviations (see Table 3).

The hierarchy of tests outlined above was performed, and the results are shown in Table 4. The LISREL syntax, with explanatory comments, is presented in Tables 5 and 6. The same basic syntax (see Table 5) was used to estimate every model, with model-specific constraints imposed in the Group 2 model line (see Table 6).

The test for form invariance, H(1), indicated that the same factor form could be applied to both groups. Even though the χ^2 value of 470.895 with 102 degrees of freedom was highly significant (indicating that one should reject the null hypothesis that the model fits the data), it was disregarded owing to the statistic's well-known sensitivity to sample size (in this case, N=1,371). Other indices, including Bentler's (1990) comparative fit index (CFI), Steiger's (1990) root mean square error of approximation (RMSEA), and Tucker and Lewis's (1973) nonnormed index (TLI), were used to assess model fit. Both RMSEA (.051) and CFI (.901) indicated an adequate fit whereas TLI (.872) indicated a marginally acceptable fit.

The test for factorial invariance, H(2), indicated that factor loadings were not significantly different between the two groups. Because of its sensitivity to large N, $\Delta\chi^2$ was not taken to be a reliable indicator of a significant difference in fit between H(1) and H(2) (Brannick, 1995; Kelloway, 1995). Here, as in all subsequent difference tests, three criteria were examined: ΔTLI greater than .05 (Little, 1997), a significant value of ΔRMSEA (Browne & Cudeck, 1993), and a probability of close fit (p_{close}) less than .05 (Browne & Cudeck, 1993). Based on these criteria, it was determined that H(3) indicated an overall failure of intercept invariance. Upon examining the individual constructs for intercept noninvariance, tests H(3.1) through H(3.3), it was discovered

TABLE 3

Correlations, Means, and Standard Deviations of Sample Data

	Mean	SD	X*1*	X*2*	X*3*	X*4*	X*5*	X*6*	X*7*	X*8*	X*9*	X*10*	X*11*	X*12*	SD	*Mean*	
*X*1	2.09	0.98	1	0.30	0.15	–0.07	0.07	–0.01	0.17	0.08	–0.18	–0.09	–0.07	–0.06	1.21	2.08	*X*1
*X*2	3.22	1.01	0.34	1	0.42	0.06	0.25	0.13	0.11	0.14	0.04	0.01	–0.06	–0.12	1.01	3.13	*X*2
*X*3	3.00	1.12	0.32	0.50	1	0.11	0.28	0.16	0.09	0.03	0.03	0.03	–0.03	–0.06	1.15	3.47	*X*3
*X*4	2.82	1.19	0.13	0.14	0.18	1	0.16	0.29	0.11	0.02	0.06	0.01	–0.05	–0.08	1.39	3.09	*X*4
*X*5	2.12	0.97	0.23	0.32	0.35	0.20	1	0.38	0.30	0.36	–0.06	–0.02	–0.09	–0.12	1.06	2.19	*X*5
*X*6	2.09	0.97	0.25	0.23	0.25	0.24	0.41	1	0.21	0.11	–0.04	–0.00	–0.09	–0.10	1.33	2.51	*X*6
*X*7	2.07	0.91	0.16	0.08	0.14	0.22	0.43	0.38	1	0.15	–0.06	0.01	–0.13	–0.12	1.22	2.70	*X*7
*X*8	2.24	0.95	0.12	0.18	0.22	0.07	0.46	0.19	0.26	1	0.01	–0.05	–0.18	–0.14	1.01	1.83	*X*8
*X*9	3.48	1.19	–0.12	–0.16	–0.06	–0.06	–0.23	–0.18	–0.08	–0.14	1	0.42	0.27	0.16	1.17	3.89	*X*9
*X*10	3.97	1.16	–0.09	–0.03	–0.03	–0.05	–0.06	–0.08	–0.02	–0.17	0.47	1	0.50	0.39	1.10	4.32	*X*10
*X*11	4.14	1.04	–0.12	–0.09	–0.12	–0.09	–0.10	–0.10	–0.06	–0.20	0.39	0.67	1	0.58	0.99	4.47	*X*11
*X*12	4.10	1.02	–0.10	–0.10	–0.13	–0.08	–0.16	–0.11	–0.08	–0.25	0.43	0.60	0.73	1	1.03	4.43	*X*12

NOTE: Italy (upper right half of the table); United States (lower left half of the table)

TABLE 4

Results of Tests for Equivalence of Latent Means: United States Versus Italy

H(#)	*Symbol*	χ^2	df	$(\Delta\chi^2)^1$	Δdf	*CFI*	*TLI*	TLI^1	*RMSEA*	$\Delta RMSEA^1$	p_{close}
H(1)	H_{form}	470.895	102	—	—	.901	0.872	—	0.051	—	.304
H(2)	H_{Λ_x}	484.801	111	13.906	9	.900	0.881	0.009	0.05	–0.001	.549
$H(3)^2$	$H_{\Lambda_x \nu_x}$	806.359	120	321.558	9	.841	0.825	–0.056	0.065	0.015*	.000
H(3.1)	$H_{\Lambda_x \nu_1}$	566.648	114	81.847	3	.882	0.863	–0.018	0.054	0.004	.073
H(3.2)	$H_{\Lambda_x \nu_2}$	713.613	114	228.812	3	.858	0.836	–0.045	0.062	0.012*	.000
$H(3.3)^2$	$H_{\Lambda_x \nu_3}$	495.854	114	1.053	3	.898	0.882	0.001	0.049	–0.001	.569
H(4.1)	$H_{\Lambda_x \nu_1 \kappa_1}$	578.539	115	11.891	1	.880	0.863	0.000	0.054	0.000	.054
H(4.2)	$H_{\Lambda_x \nu_2 \kappa_2}$	Cannot be calculated owing to intercept noninvariance: Rejection of H(3.2).									
H(4.3)	$H_{\Lambda_x \nu_3 \kappa_3}$	546.003	115	50.149	1	.890	0.874	–0.008	0.052	0.003	.188

NOTE: CFI = comparative fit index; TLI = Tucker and Lewis's nonnormed Index; RMSEA = root mean square error of approximation. Conclusions: Job context factor (ξ_1): U.S. mean ($\kappa_1^{(1)} = 2.090$) not significantly different from Italian mean ($\kappa_1^{(2)} = 2.082$). Work environment factor (ξ_3): U.S. mean ($\kappa_3^{(1)} = 3.476$) not significantly different from Italian mean ($\kappa_3^{(2)} = 3.892$).

1: Calculation of differences ($X = \chi^2$, TLI, or RMSEA):

$\Delta X(2) = X(2) - X(1)$ $\Delta X(3.3) = X(3.3) - X(2)$
$\Delta X(3) = X(3) - X(2)$ $\Delta X(4.1) = X(4.1) - X(3.1)$
$\Delta X(3.1) = X(3.1) - X(2)$ $\Delta X(4.2) = X(4.2) - X(3.2)$
$\Delta X(3.2) = X(3.2) - X(2)$ $\Delta X(4.3) = X(4.3) - X(3.3)$

2: H(3), H(3.2) indicate intercept noninvariance owing to:
ΔTLI ≈ 0.05 (Little, 1997)
Significant ΔRMSEA (Browne & Cudeck, 1993)
$p_{close} < .05$ (Browne & Cudeck, 1993)

TABLE 5
LISREL Syntax for Numerical Example (Model 1)

WORK ORIENTATION: USA	/* Title line for Group 1 (Grp 1)
DA NG=2 NI=12 NO=823	/* *N* Groups = 2, *N* Items = 12, *N S*s (Grp 1) = 823
LA	/* LAbels of items (*X*s) follow (Grp 1)
X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11 X12	
CM FI = A:\US_COR.TXT	/* Input correlation matrix file, U.S. sample (Note 1)
ME FI = A:\US_ME.TXT	/* Input observed means file, U.S. sample
SD FI = A:\US_SD.TXT	/* Input standard deviations file, U.S. sample
MO NX=12 NK=3 LX=FI TX=FR KA=FR	/* MOdel line
	/* *N* of items (X_i) = 12, *N* of constructs $\xi_j = 3$
	/* All factor loadings λ_{ij} (LX) FIxed
	/* All intercepts τ_i (TX) and means κ_j (KA) FRee
VA 1 LX 1 1 LX 5 2 LX 9 3	/* VAlue of 1 = λ_{11}, λ_{52}, & λ_{93} (Referents)
FR LX 2 1 LX 3 1 LX 4 1	/* FRee values for other λ_{ij} (λ_{21} through λ_{41})
FR LX 6 2 LX 7 2 LX 8 2	/* FRee values for other λ_{ij} (λ_{62} through λ_{82})
FR LX 10 3 LX 11 3 LX 12 3	/* FRee values for other λ_{ij} ($\lambda_{10,3}$ through $\lambda_{12,3}$)
FI TX 1 TX 5 TX 9	/* Intercepts τ_1, τ_5, and τ_9 FIxed to zero
LK	/* Labels of latent variables follow
CONTEXT CONTENT ENVIR	
ST .8 ALL	/* Use Starting value 0.80 for estimating ALL
	/* free parameters
PATH DIAGRAM	/* Output path diagram (optional)
OU	/* OUtput (last command line for Grp 1)
WORK ORIENTATION: ITALY	/* Title line for Group 2 (Grp 2)
DA NO=548	/* *N S*s (Grp 2)
LA	/* LAbels of items (*X*s) follow (Grp 2)
X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11 X12	
CM FI = A:\IT_COR.TXT	/* Input correlation matrix file, Italian sample
ME FI = A:\IT_ME.TXT	/* Input observed means file, Italian sample
SD FI = A:\IT_SD.TXT	/* Input standard deviations file, Italian sample
MO LX=PS TX=PS KA=FR	/* Model line (see Table 6)
LK	/* Labels of latent variables follow
CONTEXT CONTENT ENVIR	
ST .8 ALL	/* Use STarting value 0.80 for estimating ALL
	/* free parameters
OU	/* OUtput (last command line for Grp 2)

NOTE: If running a PC version of LISREL, enter the directory, path, and filename using DOS conventions. A data file must be in ASCII (text only) format.

that the first construct (job context) and the third construct (work environment) were invariant, but the second construct (job content) was not.

Test H(3.2) proved that the latent means of the job content variable could not be meaningfully compared across the two cultures. The means of the job

TABLE 6

LISREL Group 2 Model Lines for Testing Hypotheses

Hypothesis	*Symbol*	*Syntax (replaces Group 2 MOdel line: see Table 5)*	*Interpretation*
H(1)Form invariance	H_{form}	LX=PS TX=PS KA=FR	Current (Grp 2) factor loadings λ_{ij} (LX) have Pattern Similar to Grp 1. Current (Grp 2) intercepts τ_i (TX) have Pattern Similar to Grp 1. Current (Grp 2) latent means κ_j (KA) FRee.
H(2) Factorial invariance	H_{Λ_x}	LX=IN TX=PS KA=FR	Current (Grp 2) factor loadings λ_{ij} (LX) are INvariant (same as Grp 1). TX, KA specifications same as H(1).
H(3) Intercept invariance	$H_{\Lambda_x \nu_x}$	LX=IN TX=IN KA=FR	Current (Grp 2) intercepts τ_i (TX) are INvariant (same as Grp 1). LX, KA specifications same as H(2).
H(3.1) Intercept invariance Construct 1	$H_{\Lambda_x \nu_1}$	LX=IN TX=PS KA=FR EQ TX 1 2 TX 2 EQ TX 1 3 TX 3 EQ TX 1 4 TX 4	LX, TX, KA specifications same as H(2), plus: EQuality of Grp 1 τ_2 (TX 1 2) and current (Grp 2) τ_2 (TX 2), etc. (see Note).
H(3.2) Intercept invariance Construct 2	$H_{\Lambda_x \nu_2}$	LX=IN TX=PS KA=FR EQ TX 1 6 TX 6 EQ TX 1 7 TX 7 EQ TX 1 8 TX 8	LX, TX, KA specifications same as H(2), plus: EQuality of Grp 1 τ_6 (TX 1 6) and current (Grp 2) τ_6 (TX 6), etc. (see Note).

(continued)

TABLE 6 Continued

Hypothesis	*Symbol*	*Syntax (replaces Group 2 MOdel line: see Table 5)*	*Interpretation*
H(3.3) Intercept invariance Construct 3	$H_{\Lambda_x \nu_3}$	LX=IN TX=PS KA=FR EQ TX 1 10 TX 10 EQ TX 1 11 TX 11 EQ TX 1 12 TX 12	LX, TX, KA specifications same as H(2), plus: EQuality of Grp 1 τ_{10} (TX 1 10) and current (Grp 2) τ_{10} (TX 10), etc. (see Note).
H(4.1) Equivalent factor means, Construct 1	$H_{\Lambda_x \nu_1 \kappa_1}$	LX=IN TX=PS KA=FR EQ TX 1 2 TX 2 EQ TX 1 3 TX 3 EQ TX 1 4 TX 4 EQ KA 1 1 KA 1	LX, TX, KA specifications same as H(3.1), plus: EQuality of Grp 1 mean κ_1 (KA 1 1) and current (Grp2) mean κ_1 (KA 1).
H(4.3) Equivalent factor means, Construct 3	$H_{\Lambda_x \nu_3 \kappa_3}$	LX=IN TX=PS KA=FR EQ TX 1 10 TX 10 EQ TX 1 11 TX 11 EQ TX 1 12 TX 12 EQ KA 1 3 KA 3	LX, TX, KA specifications same as H(3.3), plus: EQuality of Grp 1 mean κ_3 (KA 1 3) and current (Grp2) mean κ_3 (KA 3).

NOTE: The syntax EQ TX 1 1 TX 1 would be redundant in H(3.1), since τ_1 is set equal to unity in both groups for the purposes of model identification. (Likewise for EQ TX 1 5 TX 5 [H(3.2)] and EQ TX 1 9 TX 9 [H(3.3)].)

context and work environment variables could be tested. However, tests H(4.1) and H(4.3) indicated no significant difference. In other words, the U.S. job context mean, 2.09, was not significantly different from the Italian job context mean of 2.08. Likewise, the U.S. work environment mean of 3.48 was not significantly different from the Italian work environment mean of 3.89. If a researcher had hypothesized a culturally based difference in either of these two variables, he or she would have been disappointed; however, he or she could take comfort from the knowledge that failure to reject the null hypothesis was not due to measurement artifacts, but rather to factual similarities between the ways that U.S. and Italian workers evaluate the quality of their jobs.

DISCUSSION

The results of the numerical example demonstrated that the American and Italian samples were invariant with respect to factor form and ERS (factor loadings). The construct of job content, however, was noninvariant with respect to ARS (intercepts). Therefore, the job content construct could not be compared across cultures on the basis of latent means.

A number of otherwise meritorious studies have not taken the issue of measurement invariance fully into account. Some (e.g., Riordan & Vandenberg, 1994) did not constrain intercepts to be invariant across groups before comparing latent means. As a result, the comparisons may have resulted in invalid inferences. Other studies have assumed ARS equivalence without testing for it (e.g., Byrne et al., 1989; Jöreskog & Sörbom, 1993; Reise et al., 1993). In particular, Byrne et al.'s study (1989) constrained the intercepts without determining whether the constraint was justified. The model having constrained intercepts displayed significantly worse fit than the model without this constraint, indicating a significant difference in ARS between groups. It is therefore possible that the reported difference in latent means may be artifacts of the difference in ARS.

As mentioned above, measures that fail the test for factorial invariance (i.e. display noninvariant ERS), may be dealt with in several ways (Poortinga, 1989). The most conservative strategy is to preclude comparisons. Because factorial invariance indicates that the conceptual frameworks of the constructs are different across cultures, a direct comparison of the factor means may be misleading. The second strategy is to eliminate noninvariant items (e.g., Dumka, Stoerzinger, Jackson, & Roosa, 1996). Though this strategy will result in an equivalent set of items representing the construct, one should pay close attention to construct validity, especially when there are only a few items to begin with (e.g., Janssens et al., 1995). The most commonly used strategy for dealing with noninvariant factor loadings is to rely on partial factorial invariance, (PFI) (e.g., Byrne, 1993; Reise et al., 1993; Riordan & Vandenberg, 1994), where the factor loadings of the invariant items are constrained to be equal across groups, and loadings of the noninvariant items are allowed to assume different values. When comparing factor means, the intercepts of the factorially noninvariant items are also allowed to vary across groups. It should be noted, however, that allowing the intercepts to vary automatically excludes the noninvariant items from the estimation of latent means. In addition, excluding items having between-group differences in ERS (e.g., Groups 1 and 2, see Figure 5) does not automatically screen out items having between-group differences in ARS (Groups 3 and 4, see Figure 5). Those items not excluded under the doctrine of PFI should be tested for intercept invariance. If this test fails, the researcher is left in a quandary, for there is (at present) no infallible, unambiguous method for identifying items that are invariant in their factor loadings but noninvariant in their intercepts.

Finally, researchers ought to treat noninvariant items as valuable information concerning cross-cultural differences and try to interpret the processes giving rise to the noninvariance. For example, one may treat noninvariant factor loadings as dependent variables and attempt to identify cultural variables that explain their variations.

Cunningham et al. (1977) recommend performing ipsative rescaling as a solution to the problem arising from differences in ERS. This is done by calculating standard scores for each participant; that is, by subtracting the mean of each participant's responses from his or her separate responses, then dividing by the standard deviation of his or her responses. The difficulty with this procedure is that any particular participant's standard deviation is in general different from those of other participants. Therefore, the transformed data are participant specific, which makes participant responses incommensurable (Horton, 1974; Stewart, 1981). In addition, ipsative rescaling changes the meaning of the original responses from " the respondent's evaluation of each item" to "the respondent's evaluation of each item relative to the other items" (Gurwitz, 1987). This not only makes interpretations difficult, but also renders the results of factor analyses invalid by imposing spurious correlations among the items (Baron, 1996; Closs, 1996). Finally, there is evidence that ipsative rescaling of data is only effective in controlling response styles when there is a large number of items (over 30) having low interitem correlations (Baron, 1996).

To support our discussion of invariance testing, we have provided examples of LISREL syntax together with detailed explanatory comments. A similar sample of syntax is provided by Little (1997) in the context of an analysis of covariance (ANCOVA) model. Our treatment is complementary because it is presented in the context of an analysis of variance (ANOVA) model.

The approach presented in this article has two limitations. First, the procedures presented here cannot tell a researcher whether important indicators of a construct have been omitted from a scale (form invariance). In addition, there are no significance tests for differences in fit indices, except for those having known sampling distributions based on the chi-squared statistic, which is sensitive to sample size. The sampling distributions of other indices (e.g., TLI) are unknown. Definitive results in this area would eliminate a major source of ambiguity encountered whenever one attempts to compare the fit statistics of competing structural models.

The aim of this article is to acquaint cross-cultural researchers with the varieties of measurement noninvariance by relating them to the familiar concepts of ERS and ARS. This article also aims to present in some detail a methodologically sound approach to the problem of noninvariance. We believe it is important that all cross-cultural researchers acquire both a theoretical understanding of these issues and a practical ability to address them using LISREL or some other SEM software. Failing in this, valid inferences concerning cross-cultural differences and similarities will be hard to discover and equally difficult to recognize once discovered.

NOTES

1. A detailed mathematical treatment of measurement invariance and factorial invariance issues can be found in Meredith (1993).

2. Although items having noninvariant factor loadings can be identified and dealt with, the situation is more ambiguous with respect to noninvariant intercepts. Unlike factor loadings, the item intercepts associated with each construct are defined in terms of the dependent variable of interest (i.e., the mean of the construct itself), and vice versa. There is no way of determining which items should be used without knowing a priori of the values of the latent means and whether they are equivalent across groups.

REFERENCES

Adler, N. J. (1984). Understanding the ways of understanding: Cross-cultural management methodology reviewed. *Advances in International Comparative Management*, *1*, 31-67.

Baron, H. (1996). Strengths and limitations of ipsative measurement. *Journal of Occupational and Organizational Psychology*, *69*, 49-56.

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

Berry, J. W., Poortinga, Y. H., Segall, M. H., & Dasen, P. R. (1992). *Cross-cultural psychology: Research and applications* (Chap. 9). Cambridge University Press.

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

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

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Newbury Park, CA: Sage.

Buss, A. R., & Royce, J. R. (1975). Detecting cross-cultural commonalties and differences: Intergroup factor analysis. *Psychological Bulletin*, *82*, 128-136.

Byrne, B. M. (1993). The Maslach Burnout Inventory: Testing for factorial validity and invariance across elementary, intermediate and secondary teachers. *Journal of Occupational and Organizational Psychology*, *66*, 197-212.

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.

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

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

Closs, S. J. (1996). On the factoring and interpretation of ipsative data. *Journal of Occupational and Organizational Psychology*, *69*, 41-47.

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

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

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.

England, G. W., & Harpaz, I. (1983). Some methodological and analytic considerations in cross-national comparative research. *Journal of International Business Studies*, *14*, 49-59.

Greenleaf, E. A. (1992). Improving rating scale measures by detecting and correcting bias components in some response styles. *Journal of Marketing Research*, *29*, 176-188.

Guilford, J. P. (1954). *Psychometric Methods*. New York: McGraw-Hill.

Guptara, P., Murray, K., Razak, B., & Sheehan, T. (1990). The art of training abroad. *Training & Development Journal*, *44*, 13-18.

Gurwitz, P. M. (1987). Ipsative rescaling: An answer to the response set problem in segmentation analysis. *Journal of Advertising Research*, *27*, 37-42.

Horton, R. L. (1974). The Edwards Personal Preference Schedule and consumer personality research. *Journal of Marketing Research*, *11*, 335-337.

Hsieh, Yu-wei (1967). Filial piety and Chinese society. In C. A. Moore (Ed.), *The Chinese mind: Essentials of Chinese philosophy and culture* (pp. 165-187). Honolulu: University of Hawaii Press.

Hui, C. H., & Triandis, H. C. (1985). Measurement in cross-cultural psychology: A review and comparison of strategies. *Journal of Cross-Cultural Psychology, 16*, 131-152.

Hui, C. H., & Triandis, H. C. (1989). Effects of culture and response format on extreme response style. *Journal of Cross-Cultural Psychology*, *20*, 296-309.

International Social Survey Program. (1989). *International Social Science Program: Work Orientations, 1989* [Computer file]. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [Distributors].

Irvine, S. H. (1969). Contributions of ability and attainment testing in Africa to a general theory of intellect. *Journal of Biosocial Science*, *1*, 91-102.

Irvine, S. H., & Carroll, W. K. (1980). Testing assessment across cultures: Issues in methodology and theory. In H. C. Triandis and J. W. Berry (Eds.), *Handbook of cross-cultural psychology* (Vol. 2). Boston: Allyn & Bacon.

Janssens, M., Brett, J. M., & Smith, F. J. (1995). Confirmatory cross-cultural research: Testing the viability of a corporation-wide safety policy. *Academy of Management Journal*, *38*, 364-382.

Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8: Structural equation modeling with the SIMPLIS command language.* Chicago: Scientific Software International.

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

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

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

Marín, G., Gamba, R. J., & Marín, B. V. (1992). Extreme response style and acquiescence among Hispanics: The role of acculturation and education. *Journal of Cross-Cultural Psychology*, *23*, 498-509.

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.

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

Moorman, R. H., & Podsakoff, P. M. (1992). A meta-analytic review and empirical test of the potential confounding effects of social desirability response sets in organizational behaviour research. *Journal of Occupational & Organizational Psychology*, *65*, 131-149.

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

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

Peterson, R. A., & Wilson, W. R. (1992). Measuring customer satisfaction: Fact and artifact. *Journal of the Academy of Marketing Science*, *20*, 61-71.

Poortinga, Y. P. (1989). Equivalence of cross-cultural data: An overview of basic issues. *International Journal of Psychology*, *24*, 737-756.

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.

Rensvold, R. B., & Cheung, G. W. (1998). Testing measurement models for factorial invariance: A systematic approach. *Educational and Psychological Measurement*, *58*, 1017-1034.

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

Schaninger, C. M., & Buss, W. C. (1986). Removing response-style effects in attribute-determinance ratings to identify market segments. *Journal of Business Research*, *14*, 237-252.

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.

Stewart, D. W. (1981). The application and misapplication of factor analysis in marketing research. *Journal of Marketing Research*, *18*, 51-62.

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

Suzuki, S., & Rancer, A. S. (1994). Argumentativeness and verbal aggressiveness: Testing for conceptual and measurement equivalence across cultures. *Communication Monographs*, *61*, 256-279.

Triandis, H. C. (1994). Cross-cultural industrial and organizational psychology. In H. C. Triandis et al. (Eds.), *Handbook of industrial and organizational psychology* (2nd ed., Vol. 4, pp. 103-172). CA: Consulting Psychologist Press.

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

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

van de Vijver, F.J.R., & Poortinga, Y. H. (1982). Cross-cultural generalization and universality. *Journal of Cross-Cultural Psychology*, *13*, 387-408.

van de Vijver, F.J.R., & Poortinga, Y.H. (1997). Towards an integrated analysis of bias in cross-cultural research. *European Journal of Psychological Assessment*, *13*, 21-29.

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

Gordon W. Cheung, Ph.D., is an associate professor in the Department of Management, the Chinese University of Hong Kong. His main research interests are structural equations modeling and cross-cultural research methods.

Roger B. Rensvold, Ph.D., is associate professor in the Department of Management, Faculty of Business at the City University of Hong Kong. His main research interests are organizational feedback and quantitative techniques for cross-cultural research.