Journal of Management
1999, Vol. 25, No. 1, 1-27

Testing Factorial Invariance across Groups: A Reconceptualization and Proposed New Method

Gordon W. Cheung
The Chinese University of Hong Kong
Roger B. Rensvold
City University of Hong Kong

Many cross-cultural researchers are concerned with factorial invariance; that is, with whether or not members of different cultures associate survey items, or similar measures, with similar constructs. Researchers usually test items for factorial invariance using confirmatory factor analysis (CFA). CFA, however, poses certain problems that must be dealt with. Primary among them is standardization, the process that assigns units of measurement to the constructs (latent variables). Two standardization procedures and several minor variants have been reported in the literature, but using these procedures when testing for factorial invariance can lead to inaccurate results. In this paper we review basic theory, and propose an extension of Byrne, Shavelson, and Muthén's (1989) procedure for identifying non-invariant items. The extended procedure solves the standardization problem by performing a systematic comparison of all pairs of factor loadings across groups. A numerical example based upon a large published data set is presented to illustrate the utility of the new procedure, particularly with regard to partial factorial invariance.

This paper attempts to advance cross-cultural management research by proposing a more rigorous technique for finding out if members of different cultures ascribe the same meanings to survey items (or similar measures). When they do, then data from all groups display the same factor loadings with respect to the same underlying constructs, a condition known as *factorial invariance*. When a set of items is not factorially invariant, then several courses of action are possible. The researcher can delete the non-invariant items, utilize partial factorial

Direct all correspondence to: Gordon W. Cheung, Department of Management, The Chinese University of Hong Kong, Shatin, Hong Kong; Phone: (852) 2609 7778; Fax: (852) 2603 6840; e-mail: <GORDONC@cuhk.edu.hk>.

invariance to retain them, or interpret them as cross-cultural data in their own right. All of these remedial techniques require that the non-invariant items be correctly identified. This paper proposes an extension of Byrne et al.'s (1989) procedure for identifying non-invariant items, which has been generally accepted as the standard procedure.

There are many issues currently stimulating interest in cross-cultural studies, including the changing demographics of the American work force, the explosive growth of international markets, and the ascendancy of the multinational organization (Triandis, 1994). Increasing cooperation across cultural boundaries makes it more important to understand culturally based differences with respect to constructs such as motivation, job satisfaction, competitiveness versus cooperation, and individualism versus collectivism. One tangible indicator of increased interest is the most recent edition of the *Handbook of Industrial and Organizational Psychology*, which includes a 869-page volume dedicated entirely to cross-cultural research. Cross-cultural and international topics are also appearing more frequently in management journals.

Cross-cultural data are frequently collected at the individual level using surveys. Before results can be compared across cultures, however, it must be shown that subjects from different cultures ascribed essentially the same meanings to the survey items. This is a special case of the more general problem of *measurement equivalence*. Psychologists and management scholars have studied measurement equivalence not only across cultures (Janssens, Brett, & Smith, 1995; Reise, Widaman, & Pugh, 1993; Riordan & Vandenberg, 1994; Windle, Iwawaki, & Lerner, 1988), but also across other groups, such as persons having different levels of academic achievement (Byrne et al., 1989), in different industries (Drasgow & Kanfer, 1985), of different gender (Byrne, 1994), and in experimental versus control groups (Pentz & Chou, 1994).

Measurement Equivalence and Factorial Invariance

Measurement equivalence exists at several different levels. If item responses (manifest variables) load on the same constructs (latent variables) across groups, and if the factor loadings are not significantly different, then *factorial invariance* is said to exist (Drasgow, 1984; Drasgow & Kanfer, 1985). This is the equivalence condition that is most frequently of interest, since it is a necessary condition for comparisons across groups (Bollen, 1989). A higher level of measurement equivalence exists if, in addition to factorial invariance, the variance-covariance matrices of the error terms are not significantly different, which indicates comparable reliability across groups (Jöreskog & Sörbom, 1989). A still higher level exists if, in addition to the above, the variances of the latent variables are not significantly different: this level is a prerequisite for comparing correlations of latent variables across groups (Jöreskog & Sörbom, 1989). Factorial invariance, however, is a necessary condition for all levels of measurement equivalence.

Since measurement equivalence in the general sense is a prerequisite for meaningful cross-cultural comparisons, a researcher should ensure that the various

versions of an instrument are identical with respect to format, instructions, and response options. The items should be translated accurately, usually by following a blind back-translation strategy. One qualified person translates the instrument from language A into language B. A second person, working without reference to the original instrument, translates it from language B back into A. The second A version is compared with the original, and the translators confer to resolve discrepancies (Brislin, Lonner, & Thorndike, 1973).

Even after reasonable precautions have been taken to prepare an equivalent instrument, it is possible that some items may have significantly different meanings for one group than for another. These differences will be reflected in the factor loadings. For example, agreeing with a statement such as, "I am a person of worth, at least as good as other people," may indicate a healthy level of self-esteem—to an American. To a Chinese, on the other hand, it may seem that agreeing to such a statement indicates a grandiose, socially unacceptable sense of self-importance. As a result, the item may have a weaker loading on the construct of "self-esteem" for Chinese subjects than for American subjects, even if the item has been accurately translated from English into Chinese.

The problem of factorial invariance poses two questions. First, is the list of items, taken as a whole, invariant across cultures? If the answer to this question is in the negative, then the second question is, which items are non-invariant? Although problems of this type have been discussed for more than thirty years (Meredith, 1964a, 1964b), tests of invariance using Confirmatory Factor Analysis (CFA) have only become common in the last ten years.

Since factorial invariance is sometimes difficult to achieve, some researchers (Byrne et al., 1989; Marsh & Hocevar, 1985) have proposed relaxing it as a prerequisite for cross-cultural comparisons, and rather relying upon *partial* factorial invariance. Under partial factorial invariance, non-invariant items are retained, and their loadings are allowed to vary when analyzing between-group differences. It is assumed that if the non-invariant items constitute only a small portion of the model, then they will not affect cross-group comparisons to any significant extent. However, researchers cannot determine whether or not partial factorial invariance is appropriate unless they can identify the non-invariant items, and determine the extent of their departure from invariance (that is, the extent to which their factor loadings actually differ across groups).

Testing for factorial invariance requires estimating a number of measurement models using CFA (as discussed in greater detail below). One essential preliminary step in estimating any CFA model is to provide *standardization*, that is, to assign units of measurement to the latent variables. There are several procedures for providing standardization; unfortunately, those in current use may produce incomplete and misleading tests of factorial invariance.

The remainder of this paper is arranged as follows. A flowchart (Figure 1) is used to describe a series of tests for factorial invariance. These tests extend the procedure proposed by Byrne et al. (1989). The proposed extensions are explained and justified. Finally, a summary of the extended procedure is shown in tabular form (Table 1), and a numerical example is presented.

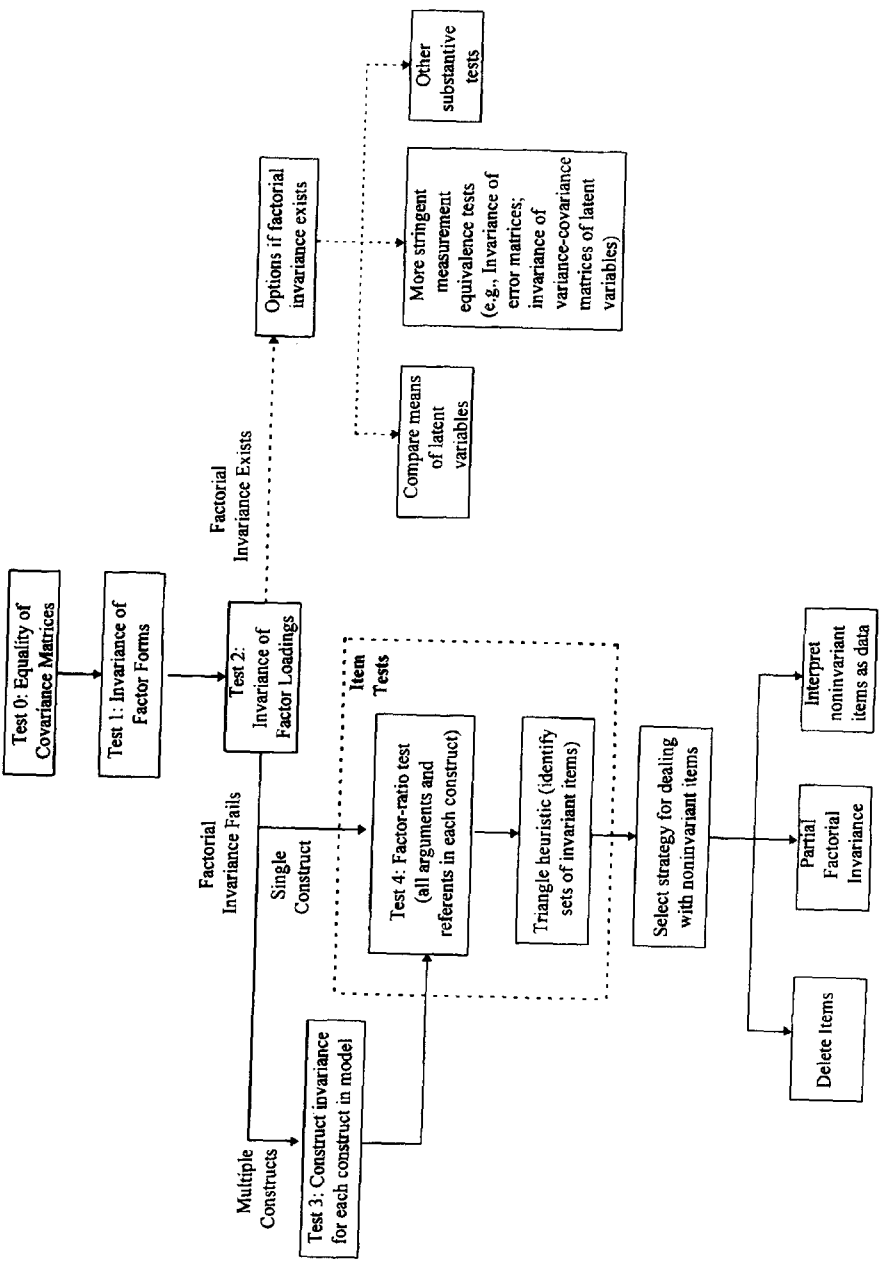

Figure 1. Tests for Factorial Invariance

Table 1. Tests for Factorial Invariance

Test	*Null Hypothesis (H_0):*	*Test Statistic(s)*	*If test statistic significant (reject H_0), then —*	*If test statistic n.s. (fail to reject H_0), then —*
0: Equality of Covariance Matrices	For all groups g: $\Sigma^{(1)} = \Sigma^{(2)} \ldots = \Sigma^{(g)}$	χ^2	Go to test 1	Factorial invariance exists. Continue with research questions.
1: Invariance of Factor Form	For all groups g: $\Lambda_{form}^{(1)} = \Lambda_{form}^{(1)} = \ldots = \Lambda_{form}^{(g)}$	χ^2_{uncon}, CFI, TLI, other fit indices.	STOP. Inadequate baseline model.	Go to test 2.
2: Factorial invariance	For all i, j, g in the model: $\lambda_{ij}^{(1)} = \lambda_{ij}^{(2)} = \ldots = \lambda_{ij}^{(g)}$, or $\Lambda_x^{(1)} = \Lambda_x^{(2)} = \ldots = \Lambda_x^{(g)}$	$\Delta\chi^2 = \chi^2_{con} - \chi^2_{uncon}$, changes in other fit indices.	Invariance fails. If there are multiple constructs in the model, continue with Test 3; if there is a single construct, go to Test 4.	Factorial invariance exists. Continue with research questions (or perform more stringent measurement equivalence tests, if required.)
3: Factorial invariance (for each construct)	For all i, g within each construct j: $\lambda_{ij}^{(1)} = \lambda_{ij}^{(2)} = \ldots = \lambda_{ij}^{(g)}$,	$\Delta\chi^2 = \chi^2_{con} - \chi^2_{uncon}$, changes in other fit indices.	Invariance fails for one or more items in this construct. Go to Test 4.	Continue until all constructs have been tested.
4: Factor-ratio test (for each nonequivalent construct).	For all i, g within each nonequivalent construct j: $(i \neq i')$ $\frac{\lambda_{ij}^{(1)}}{\lambda_{i'j}^{(1)}} = \frac{\lambda_{ij}^{(2)}}{\lambda_{i'j}^{(2)}} = \ldots = \frac{\lambda_{ij}^{(g)}}{\lambda_{i'j}^{(g)}}$	$\Delta\chi^2 = \chi^2_{con} - \chi^2_{uncon}$, changes in other fit indices.	Invariance fails for this ratio. Record ratio and analyse along with the other ratios using the triangle heuristic.	This ratio invariant. Record the ratio and analyse along with the other ratios using the triangle heuristic.
		Triangle Heuristic: Identify invariant sets of items. Select strategy for dealing with non-invariant items. Continue with research questions.		

The Sequence of Non-invariant Tests[1]

The overall test sequence follows the tradition established by Jöreskog and Sörbom (1989: 230). First, the covariance matrices of the groups are compared[2] (Test 0, Figure 1); that is,

$$H_0: \Sigma^{(1)} = \Sigma^{(2)} = \ldots = \Sigma^{(g)},$$

for all groups g. If the null hypothesis cannot be rejected, then factorial invariance exists. This, however, is an extremely rigorous test. One would not expect groups to be invariant in this sense unless they were drawn at random from the same population, a situation having little relevance to cross-cultural studies. Therefore, published research typically does not include the results of this test.

Test 1 (Figure 1) tests the fit of a theoretically derived baseline model. The pattern of significant factor loadings between manifest and latent variables is tested for invariance across groups; that is,

$$H_0: \Lambda_{form}^{(1)} = \Lambda_{form}^{(2)} = \ldots = \Lambda_{form}^{(g)}$$

for all groups g. Factor loadings are not constrained to be equal across groups for this test. If the overall fit is not adequate with respect to appropriate statistics (χ^2, CFI, TLI, etc.), then an adequate baseline model does not exist. The results indicate that either some items load on different factors across groups, or different groups produce different numbers of factors, or both. If this situation exists then further tests are not performed; otherwise, the test process continues. The χ^2 statistic[3] associated with this unconstrained baseline model, χ^2_{uncon}, is used in subsequent tests.

Test 2 (Figure 1) compares the baseline model with a fully constrained model in which all factor loadings are required to be identical across groups; that is,

$$H_0: \Lambda_x^{(1)} = \Lambda_x^{(2)} = \ldots = \Lambda_x^{(g)}$$

for all g. The fit statistic for this model is χ^2_{con}. The fully constrained model is compared with the baseline model by calculating $\Delta\chi^2 = \chi^2_{con} - \chi^2_{uncon}$. If $\Delta\chi^2$ is not statistically significant, then factorial invariance exists. If $\Delta\chi^2$ is significant, then the unconstrained baseline model fits the data more closely than the constrained model, indicating that the constrained model could be improved by relaxing one or more of the equality constraints. Factorial invariance, therefore, does not exist.

If factorial invariance exists, then the researcher has several options to pursue based upon the substantive research questions. First, the means of latent variables can be compared (e.g., Byrne et al., 1989; Millsap & Everson, 1991; Riordan & Vandenberg, 1994). Factorial invariance is a necessary condition for comparing means and intercepts in a factor mean structure across cultural groups (Bollen,

1989). The second option is to conduct other substantive tests. For example, Singh (1995) compared path coefficients of a structural model across cultures after establishing factorial invariance. Another option is to continue with the tests of more stringent levels (increasing restrictive tests) of measurement equivalence (Jöreskog & Sörbom, 1989). If the researcher wants to compare scale reliabilities across groups, then the test for equivalence of the error variance-covariance matrices should be performed (e.g., Byrne, 1994; Drasgow & Kanfer, 1985; Marsh & Hocevar, 1985; Mullen, 1995). If the study seeks to compare correlations of latent variables across groups, then the test for the equality of the variance-covariance matrices of the latent variables should be performed (e.g., Byrne, 1994; Marsh, 1993; Jackson, Wall, Martin, & Davids, 1993; Marsh & Hocevar, 1985).

If factorial invariance does not exist, then additional tests are required to determine the sources of non-invariance. If the measurement model consists of multiple constructs, then the next step in the sequence is Test 3. If the model consists of only a single construct, then Test 3 is omitted and the procedure continues with Test 4 (below).

Test 3 (Figure 1) examines each of the constructs in the model for invariance. A separate model is estimated for each construct. In each model, the factor loadings associated with the construct are constrained to be equal across groups, while the loadings associated with the other constructs are not (Table 1). The χ^2 fit statistic of each model is used to calculate $\Delta\chi^2$, as above. If $\Delta\chi^2$ is significant, then at least one of the items within the construct is non-invariant. All non-invariant constructs are noted, and their items are examined for invariance in Test 4.

Test 4 (Figure 1) is Byrne et al.'s (1989) procedure, which makes a cross-group comparison of each of the loadings associated with each of the constructs identified under Test 3. Once again, a series of tests is performed. A separate model is estimated for each item, in which that item's loading is constrained to be equal across groups (Table 1). As in Test 3, the χ^2_{con} fit statistic of each constrained model is compared with the baseline χ^2_{uncon}. If a particular $\Delta\chi^2$ is significant, then that item is non-invariant.

It should be noted that three other procedures have been used to identify non-invariant items. The first procedure examines the factor loadings of the unconstrained model, and those having the greatest difference between groups are identified as non-invariant (e.g., Van de Vijver & Harsveld, 1994). Although straightforward, this procedure is suspect because it does not incorporate significance tests of the observed differences. The second procedure examines the significance of factor loadings; a loading is identified as non-invariant if it is significant for one group but not another (e.g., Janssens et al., 1995). Reflecting the arbitrary nature of significance levels, the procedure is problematical in cases where the significance levels are nearly equal, yet one is significant and the other is not. The third procedure utilizes modification indices (MIs) (e.g., Marsh & Hocevar, 1985; Reise et al., 1993; Riordan & Vandenberg, 1994). A large MI in the fully constrained model indicates that the constraint ought to be relaxed in order to improve the fit, and the item is, therefore, taken to be non-invariant. Unfortunately, the algorithm used to estimate MIs makes this procedure suspect. As the MI for each item is calculated in the fully constrained model, all other items in the model

are assumed to be invariant. Since restrictions on one set of factor loadings also generally affect the estimated values of other coefficients (Williams & Thomson, 1986), if the invariant assumption of all other factor loadings is not valid (for example, when one or more of the other loadings is not invariant), then the value of the MI may not be accurate.

The procedure referred to here as Test 4 (Byrne et al., 1989) has none of these theoretical difficulties. The loadings are tested for equivalence one at a time, subject to no assumptions about relative size, significance, or the non-invariance of other loadings in the model. There is, however, a problem (the *raison d'être* of this paper) associated with estimating the constrained models in Test 4. The problem has to do with the standardization requirement.

The Standardization Problem

When estimating a CFA model, the researcher must assign a unit of measurement to the construct (latent variable) by using some standardization procedure (Jöreskog & Sörbom, 1989). All such procedures embody a tacit *assumption* of invariance, even though the purpose of the procedures described above is to *test* for invariance. Problems may arise when the entity used for standardization is not actually invariant.

A simple four-item, one-construct, two-group measurement model is shown in Figure 2. Two types of standardization procedure (plus minor variants) are in general use. Type 1 standardizes the construct variance across groups (i.e., $\phi_{11}^{(1)} = \phi_{11}^{(2)} = 1$). Type 2 involves selecting an item (other than the one being tested for invariance) and setting its factor loading to unity. For example, if X_1 were being tested, then one could select X_2 as the standardization item and set the constraint $\lambda_{21}^{(1)} = \lambda_{21}^{(2)} = 1$. The construct ξ_1 would then be estimated utilizing X_2's unit of measurement.

Type 1 standardization (e.g., Bandalos, 1993) is the less common of the two. Standardizing the construct defines the metric applied to the estimated factor loadings of that group (MacCallum & Tucker, 1991). If the true variances of the constructs are not equal across groups, then tests for factorial invariance may be biased because the factor loadings for each group are using different metrics. For example, setting $\lambda_{11}^{(1)} = \lambda_{11}^{(2)}$ when the data indicates otherwise may result in erroneously rejecting the hypothesis $\phi_{11}^{(1)} = \phi_{11}^{(2)}$ (or one of the other three hypotheses) when it is true. In addition, Type 1 standardization results in a test for *strict* factorial invariance (Meredith, 1993), in which both the factor loadings and the variances of the constructs are invariant (e.g., Aiken, Stein, & Bentler, 1994; Byrne, 1994). It is unnecessarily stringent for many investigations, and may also result in an ambiguous test for factorial invariance (Appendix A).

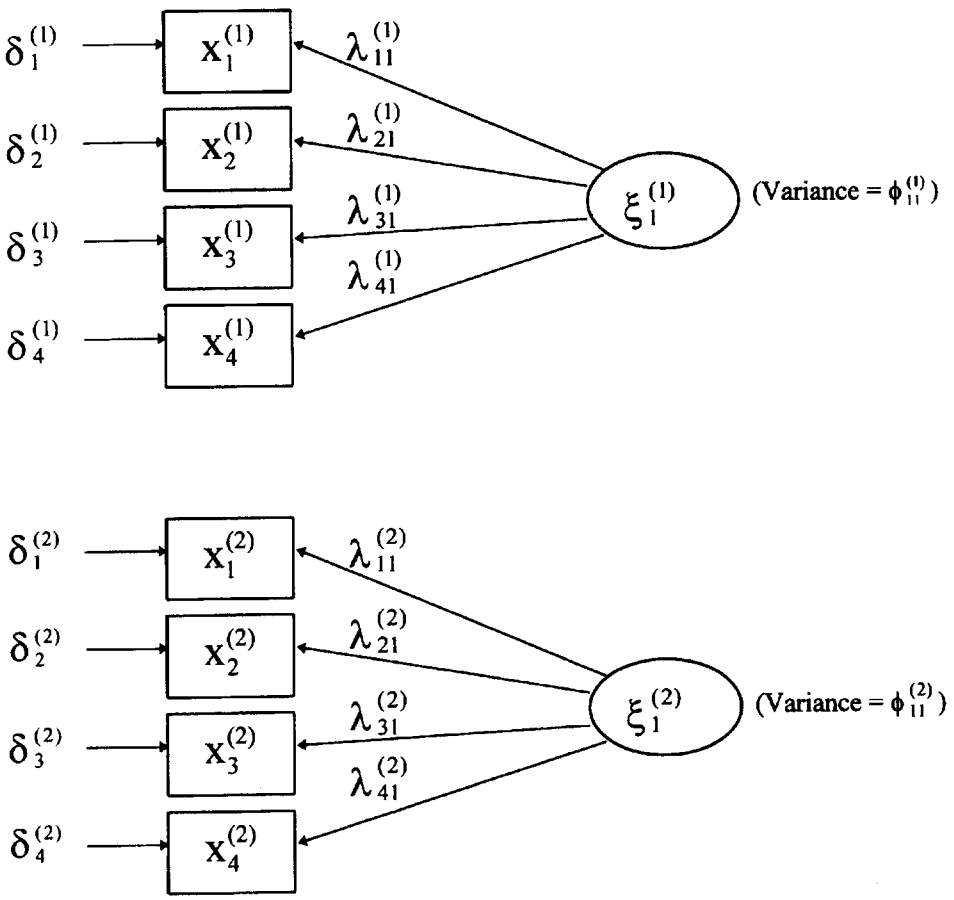

Figure 2. Multisample Factorial Model (Two Groups)

Type 2 standardization is the most common procedure (Riordan & Vandenberg, 1994; Smith, Tisak, Bauman, & Green, 1991; Van de Vijver & Harsveld, 1994). One selects an item other than the item being tested for invariance, and sets its factor loading equal to unity across groups. For the remainder of the paper, the selected item is referred to as the *referent.* To make clear the distinction between the referent and the item being tested, we refer to the latter as the *argument.*

By setting the factor loading of the referent equal to unity across groups, Type 2 standardization tacitly assumes that the true factor loadings of the referent *are* equal across groups. If this assumption is incorrect, it can lead to inaccurate estimates of other model parameters (Bollen, 1989), and comparisons of factor loadings across groups may be biased. This argument is presented in greater detail in Appendix B.

A variation of Type 2 standardization, referred to here as Type 2a, (Drasgow & Kanfer, 1985) begins by setting the variance of one construct to unity. Then the

whole sample, without regard to group membership, is used to estimate factor loadings. The item having the largest factor loading is used as the referent. Although the loading of the referent is not constrained to equal 1.0 across groups, it is still required to assume a same value in all groups. Therefore, Type 2 and Type 2a standardization suffer from the same problem, that they assume the relation between the referent and latent variable is the same across groups.

A second variation of Type 2 standardization, here referred to as Type 2b, is suggested by Reise et al. (1993). This technique standardizes the variance of one construct, and then constrains the factor loadings of an item to be equal across groups. Under this procedure the metric for the first group is specified by standardizing the construct, and the same metric is extended to the second group by way of the constrained item. Like Type 2, Type 2b standardization assumes that the referent is invariant across groups. In fact, the statistics for this procedure are identical to Type 2 statistics, since they both make the same assumption regarding the equality of the referent loadings, and differ only in that Type 2b standardization does not set them to unity.

In summary, Type 1 standardization is overly restrictive for most applications, and assumes that the variances of the constructs are equal across groups. There are three varieties of Type 2 standardization, each of which makes an *a priori* assumption about the invariance of the *true* factor loadings of the referent across groups. Unfortunately, this is an untestable assumption since only the invariance of the *ratios* of factor loadings can be tested for invariance across groups (Bielby, 1986; Williams & Thomson, 1986). We propose a solution based upon Type 2 standardization, which we refer to as the factor-ratio test. Instead of using only one item as a referent, the procedure uses *all* of them, following an iterative scheme.

The Factor-Ratio Test

In Test 4, a series of constrained models are estimated, one for each item i in each non-invariant construct j. It follows that there is not just one, but rather two constraints associated with every model of this type; i.e.,

$$\lambda_{ij}^{(1)} = \lambda_{ij}^{(2)} = \ldots = \lambda_{ij}^{(g)} \qquad \text{(the test constraint)}$$

and $$\lambda_{i'j}^{(1)} = \lambda_{i'j}^{(2)} = \ldots = \lambda_{i'j}^{(g)} = 1 \qquad \text{(the standardization constraint),}$$

where $i \neq i'$.

The standardization constraint assumes that referent $X_{i'}$ is invariant across groups. If it is not, then using it as the referent may bias the test of the argument X_i. In order to test all for invariance, it is necessary to construct and test a model for *each combination* of X_i and $X_{i'}$, subject to the constraints above. This is an extension of Byrne et al.'s (1989) procedure, but it is a far-reaching extension. Details supporting the recommendation are given in Appendix B, where it is shown that each model produces a test for the null hypothesis

$$\frac{\lambda_{ij}^{(1)}}{\lambda_{i'j}^{(1)}} = \frac{\lambda_{ij}^{(2)}}{\lambda_{i'j}^{(2)}} = \ldots = \frac{\lambda_{ij}^{(g)}}{\lambda_{i'j}^{(g)}}$$

for all groups g. The systematic examination of all combinations of referents and arguments, across all groups, is the factor-ratio test.

If there are n items to be tested, then $n(n-1)/2$ tests are required. It is convenient to arrange the results of these tests in matrix form, with numbered rows corresponding to the arguments X_i and columns representing the referents $X_{i'}$. As an example, consider the construct ξ_1 shown in Figure 2. Assume this construct is non-invariant, as determined by Test 3. Testing the four items for invariance requires six separate tests, the results of which are shown in Exhibit 1a. Table entry C* is the result of testing argument X_3 for invariance using referent X_2 (and also the result of testing argument X_2 using referent X_3: see Appendix B). As before, C* is the value of $\Delta\chi^2 = \chi^2_{con} - \chi^2_{uncon}$. The asterisk indicates that C* is statistically significant; that is, X_3 is not invariant *when tested using* X_2 as the referent.

The Triangle Heuristic

It is proposed that an item only be considered invariant if it belongs to an *invariant set*. An item belongs to such a set if it is invariant when tested using *all other* members of the set as referents. A systematic procedure can be applied to identify invariant sets of items, beginning with the test data in the tabular form as shown above.

It is possible to rearrange the table without changing the meaning of its entries by swapping rows while simultaneously swapping the corresponding columns. For example, it is permissible to rewrite the row 1 entries into row 2 (and vice versa) if one also rewrites the column 1 entries into column 2 (and vice versa). This procedure can be used to identify an invariant set of items by using a "triangle heuristic." The heuristic consists of systematically swapping rows and columns with the goal

Exhibit 1. Results of Factor-Ratio Tests (Four Items)

a

Arguments	Referents X_1	X_2	X_3	X_4
X_1	—	A	B	D
X_2	A	—	C*	E
X_3	B	C*	—	F*
X_4	D	E	F*	—

b

Arguments	Referents X_1	X_2	X_4	X_3
X_1	—	A	D	B
X_2	A	—	E	C*
X_4	D	E	—	F*
X_3	B	C*	F*	—

Exhibit 2. Triangle Heuristic—Two Equivalent Sets

a

Invariant Set 1

Arguments	Referents X_1	X_2	X_3	X_4
X_1	—	A	B	D
X_2	A	—	C	E*
X_3	B	C	—	F
X_4	D	E*	F	—

b

Invariant Set 2

Arguments	Referents X_1	X_3	X_4	X_2
X_1	—	B	D	A
X_3	B	—	F	C
X_4	D	F	—	E*
X_2	A	C	E*	—

of producing the largest possible closed triangular array of non-significant entries below the diagonal, with the apex of the triangle in the row 2, column 1 position. Once this has been achieved, the items defining the rows and columns of the triangle (including the diagonal) are an invariant set.

Beginning with the data shown in Exhibit 1, it is possible to swap rows 3 and 4, and columns 3 and 4. This results in Exhibit 1b, with the triangle entries shaded. The invariant set of items for this construct are items 1, 2, and 4. The non-invariant item is item 3.

A researcher may identify more than one invariant set within a set of items that initially represented only one construct. The matrix on the left side of Exhibit 2 (Exhibit 2a) is already in triangular form, identifying an invariant set consisting of items 1, 2, and 3. The table can be rearranged into another, different triangular form (Exhibit 2b) consisting of items 1, 3, and 4.

This result derives from a fundamental fact. The item responses *per se* do not define a construct, but rather the relationships among them; i.e., their covariance structure. The between-group invariance constraint may have the effect of "pulling apart" the covariance structure. Instead of one set of covarying items, there may be two (or more) sets of items that are invariant across groups, but which represent slightly different ideas. The decision about which (if any) of the invariant sets should be used in the context of a particular research project must be made in light of substantive issues and underlying theory. The items used must have construct validity, as well as factorial invariance. The example below illustrates this point.

A Numerical Example

The following example uses the "Work Orientations" data set published by the International Social Survey Program (ISSP, 1989), an international project begun in 1984. The overall sample consists of 14,733 persons 18 years and older from West Germany, Great Britain, Northern Ireland, Austria (14 years and older), Norway, Hungary, the Netherlands (16 years and older), Italy (the Italian population), and the United States (noninstitutionalized English-speaking persons only).

Table 2. Perceived Aspects of Job Quality

Item	*Item Content*	*Responses*
Construct 1 (ξ_1): Quality of Work Environment (WE)		
X_1	Hard physical work	1 (Always) to 5 (Never)
X_2	Work in dangerous conditions	1 (Always) to 5 (Never)
X_3	Unhealthy conditions	1 (Always) to 5 (Never)
X_4	Physically unpleasant conditions	1 (Always) to 5 (Never)
Construct 2 (ξ_2): Quality of Job Context (JX)		
X_5	Secure job	1 (Strongly agree) to 5 (Strongly disagree)
X_6	High income	1 (Strongly agree) to 5 (Strongly disagree)
X_7	Good opportunities	1 (Strongly agree) to 5 (Strongly disagree)
X_8	Flexible working hours	1 (Strongly agree) to 5 (Strongly disagree)
Construct 3 (ξ_3): Quality of Job Content (JC)		
X_9	Interesting job	1 (Strongly agree) to 5 (Strongly disagree)
X_{10}	Independent work	1 (Strongly agree) to 5 (Strongly disagree)
X_{11}	Help other people	1 (Strongly agree) to 5 (Strongly disagree)
X_{12}	Bored at work (reverse scored)	1 (Never) to 5 (Always)

This analysis used two groups, West Germans ($n = 591$) and Americans ($n = 823$). The model to be tested for equivalence consists of three correlated constructs measuring different perceived aspects of work: quality of work environment (WE: ζ_1), quality of job context (JX: ξ_2), and quality of job content (JC: ξ_3). Each construct is represented by four items, each scored on a 1 to 5 Likert scale (Table 2). The correlations, means and standard deviations for the two samples are shown in Table 3. The sequence of non-invariant tests (Figure 1) were performed based on covariance matrices.

Because it was assumed there would be significant differences in the covariance matrices of the two groups, Test 0 (Figure 1) was not performed. Test 1 (Figure 1) examined the fit of an unconstrained baseline model. The fit is satisfactory (Table 4: $\chi^2(102) = 397.00$; CFI = 0.93; TLI = 0.91).

Test 2 (Figure 1) began by estimating a fully constrained model. This model fit the data less well than the baseline model estimated in Test 1 (Table 4: $\Delta\chi^2 = 68.63$, $p < .0001$). The significance level was adjusted to control for experiment-wise Type 1 error.[4] This finding indicated that the model as a whole did not possess factorial invariance.

Test 3 (Figure 1) compared the constructs across groups. Three constrained models, one for each target construct, were estimated and tested against the baseline. The factor loadings of the target construct in each model were constrained to be invariant across groups; factor loadings of other constructs were not constrained. Table 4 presents the results of Test 3.1 (for construct WE), Test 3.2 (JC) and Test 3.3 (JX). Factorial invariance did not exist for two of the constructs, since their fit statistics differed from the baseline model at the 0.0001 level of significance: they were WE ($\Delta\chi^2 = 41.55$) and JC ($\Delta\chi^2 = 25.72$).

Table 3. Means, Standard Deviations, and Correlations[1]

					Germany											
					X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}
				M	3.74	4.21	4.09	3.96	1.85	2.97	3.15	3.54	1.99	2.04	2.61	1.68
				SD	1.22	1.03	1.09	1.11	0.87	0.95	1.06	1.35	0.90	0.94	1.16	0.84
		M	**SD**													
	X_1	3.48	1.19			0.54	0.48	0.61	–0.06	–0.09	–0.11	0.02	–0.20	–0.14	–0.00	–0.03
	X_2	3.97	1.16		0.47		0.56	0.56	0.00	0.02	–0.01	–0.04	–0.16	–0.15	–0.05	–0.10
	X_3	4.14	1.04		0.39	0.67		0.60	–0.08	–0.01	–0.06	–0.06	–0.19	–0.19	–0.05	–0.15
	X_4	4.10	1.02		0.43	0.61	0.73		–0.09	–0.06	–0.08	–0.01	–0.26	–0.18	–0.00	–0.15
	X_5	2.09	0.98		–0.12	–0.09	–0.12	–0.10		0.32	0.26	0.05	0.31	0.25	0.22	0.16
	X_6	3.22	1.01		–0.16	–0.03	–0.09	–0.10	0.34		0.43	0.18	0.30	0.30	0.18	0.15
US	X_7	3.00	1.12		–0.06	–0.03	–0.12	–0.13	0.32	0.50		0.15	0.32	0.27	0.23	0.10
	X_8	2.82	1.19		–0.06	–0.05	–0.09	–0.08	0.13	0.14	0.18		0.07	0.26	0.11	0.01
	X_9	2.12	0.97		–0.23	–0.06	–0.10	–0.16	0.23	0.32	0.35	0.20		0.57	0.43	0.33
	X_{10}	2.09	0.97		–0.18	–0.08	–0.10	–0.11	0.25	0.23	0.25	0.24	0.41		0.37	0.33
	X_{11}	2.07	0.91		–0.08	–0.02	–0.06	–0.08	0.16	0.08	0.14	0.22	0.43	0.38		0.20
	X_{12}	2.24	0.95		–0.14	–0.17	–0.20	–0.25	0.12	0.18	0.22	0.07	0.46	0.19	0.26	

Note: [1]Correlations from U.S. sample (n = 823) are shown below the diagonal; correlations from Germany sample (n = 591) are shown above the diagonal.

Table 4. Tests of Factorial Invariance, Perceived Aspects of Job Quality

Test	*Model*	χ^2	*df*	$\Delta\chi^2$	Δ *df*	*TLI*	*CFI*
1	Unconstrained baseline model	397.00	102			.91	.93
2	Fully constrained model	465.64	111	68.63**	9	.91	.92
3.1	Loadings on WE (ξ_1) constrained	438.55	105	41.55**	3	.91	.93
3.2	Loadings on JC (ξ_2) constrained	422.72	105	25.72**	3	.91	.93
3.3	Loadings on JX (ξ_3) constrained	398.25	105	1.24	3	.92	.93

Notes: WE = work environment
JX = job context
JC = job content
Models identified by fixing λ_{11}, λ_{52} and λ_{93} to 1.0.
** $p < .0001$

Test 4 (Figure 1) used Byrne et al.'s (1989) procedure, extended by the factor-ratio test, to examine the items within each non-invariant construct. A separate model was estimated for each combination of argument and referent, and tested against the baseline for a significant difference in fit. In the interests of clarity we will describe one of the factor-ratio tests in detail. The items being tested were those associated with ξ_1. The first test utilized X_1 as the referent and X_2 as the argument. In this model, λ_{11} was constrained to equal unity for both groups; that is, $\lambda_{11}^{(1)} = 1$ and $\lambda_{11}^{(2)} = 1$. Because of the choice of argument, λ_{21} was constrained to be equal across groups; that is, $\lambda_{21}^{(1)} = \lambda_{21}^{(2)}$. All other factor loadings within ξ_1 were allowed to vary. In addition, all factor loadings associated with the other two constructs, ξ_2 and ξ_3, were allowed to vary (except for λ_{52} and λ_{93} which were also fixed to unity to provide identification for the model). The resulting χ^2 statistic was compared with the χ^2 of the unconstrained baseline model estimated in Test 1. The chi-square difference (35.10) was entered in Exhibit 3a as data for the triangle heuristic. Preliminary results for the WE construct (Exhibit 3a) indicated that X_2 and X_3 were not invariant when using X_1 as the referent. Rearranging the table in accordance with the triangle heuristic (Exhibit 3b) shows that there was one invariant set associated with WE; namely, X_2, X_3, and X_4. The item X_1 was not invariant.

The factor-ratio test was also applied to the items associated with the JC construct (Exhibit 4). Preliminary results (Exhibit 4a) indicated that item X_{12} was not invariant when using X_{10} as the referent. Without rearrangement, Exhibt 4b identified an invariant set in accordance with the triangle heuristic; namely, items X_9, X_{10}, and X_{11}. Rearranging the data using the heuristic identified a second set; X_9, X_{11}, and X_{12} (Exhibit 4c).

This example highlights an important point. Factorial invariance did not exist with respect to the JC construct ($\Delta\chi^2 = 25.72$). However, testing the items using *only* X_9 as a referent would result in *no* item being identified as a source of the invariance (note the first column in Exhibit 4b, all the entries are non-significant).

Exhibit 3. Identifying Invariant Sets of Items Measure of Work Environments (WE)

a: Preliminary results from factor-ratio tests

	Referents			
Arguments	X_1	X_2	X_3	X_4
X_1	—			
X_2	35.10**	—		
X_3	29.70**	0.37	—	
X_4	11.76	11.36	9.01	—

b: Identifying invariant set X_2, X_3 and X_4 using the triangle heuristic:

	Referents			
Arguments	X_2	X_3	X_4	X_1
X_2	—			
X_3	0.37	—		
X_4	11.36	9.01	—	
X_1	35.10**	29.70**	11.76	—

Note: $^{**}p < 0.0001$.

The point can be generalized to any test of this nature. Testing for factorial invariance using only one referent (that is, *failing* to use the factor-ratio method demonstrated above) may identify no non-invariant items, even if the overall construct is non-invariant.

The results also give the researcher options about how to deal with invariance failure. If he or she wishes to render the JC construct factorially invariant by deleting an item (not usually a good idea, see the discussion below), then the choice is between X_{10} and X_{12}. If X_{10} is deleted, then the "quality of job content" construct loses an element relating to job independence. If X_{12} is deleted, then the construct loses an element relating to lack of boredom. The choice must be made based upon the theory and the substantive issues motivating the research.

The doctrine of partial factorial invariance (PFI) allows a researcher to argue that all four of the JC items should be used. The question then becomes, which of the two invariant sets should be used. The researcher could make that decision based upon its implications for cross-group comparisons. If job independence has substantively less interest than lack of boredom, then under PFI the loadings associated with X_9, X_{11}, and X_{12} are constrained to be invariant across groups, while the X_{10} loading is allowed to vary. If, on the other hand, lack of boredom has less inter-

Exhibit 4. Identifying Sets of Factorially Invariant Items Measure of Job Content (JC)

a: Preliminary results from factor-ratio tests

Arguments	Referents X_9	X_{10}	X_{11}	X_{12}
X_9	—			
X_{10}	12.07	—		
X_{11}	6.97	0.18	—	
X_{12}	3.78	20.20**	13.63	—

b: Identifying invariant set X_9, X_{10} and X_{11} using the triangle heuristic:

Arguments	Referents X_9	X_{10}	X_{11}	X_{12}
X_9	—			
X_{10}	12.07	—		
X_{11}	6.97	0.18	—	
X_{12}	3.78	20.20**	13.63	—

c: Identifying alternative invariant set X_9, X_{11} and X_{12} using the triangle heuristic:

Arguments	Referents X_9	X_{11}	X_{12}	X_{10}
X_9	—			
X_{11}	6.97	—		
X_{12}	3.78	13.63	—	
X_{10}	12.07	0.18	20.20**	—

Note: $^{**}p < 0.0001$.

est, the loadings associated with X_9, X_{10}, and X_{11} can be constrained, and the X_{12} loading allowed to vary.

Discussion

Factorial invariance is a critical issue in cross-cultural studies. Without it, interpretations of differences across cultures are problematical, since it is not

established that members of different groups are using the same conceptual frames of reference when responding to survey items (Riordan & Vandenberg, 1994). Therefore, a series of tests for factorial invariance should be a prerequisite for any cross-cultural comparison.

Byrne et al. (1989) provided a powerful procedure for testing factorial invariance. The procedure, however, requires a choice of referent to identify the model. This choice, while seemingly incidental, is actually crucial to obtaining accurate results. Any referent can be used when testing for full factorial invariance, but the use of different referents may lead to different results when testing for invariance at the item level. Hence, the choice of referent affects identification of the non-invariant items, which in turn affects subsequent comparisons of factor means and of structural models across groups (Millsap & Everson, 1991). Reise et al. (1993) recommend that the choice of referents not be made arbitrarily, but rather be supported by past studies. However, many researchers are not aware of this recommendation and choose referents without offering any rationale for their choices.[5]

The recommendation presented here employs an extension of Byrne et al.'s (1989) procedure that systematically examines all non-redundant combinations of referents and arguments, leading to identification of invariant *sets* of items. The new procedure may identify more than one invariant set related to each construct. When this occurs, the researcher must decide which set to use, based upon both substantive and theoretical considerations. This outcome emphasizes an important and extremely fundamental point. Items are factorially invariant to the extent that the items display common *patterns* of responses across groups.

Items that are not factorially invariant may be dealt with in several ways (Poortinga, 1989). First, the items may be simply eliminated from the scale. This option should be considered with caution, however, since it tends to be atheoretical. Theories are intended to provide frameworks that explain as many phenomena as possible, subject to the antithetical demands of usefulness and parsimony. The factor model being tested for invariance is supposedly based upon theory, which in turn is based upon the cumulative experience of previous investigators. Arbitrarily trimming items from the model may increase its usefulness in one particular instance, but it is at odds with the principle of parsimony, since the logical endpoint of this sort of exercise is to produce as many models as there are data sets. From a practical point of view, dropping items may pose a problem if there are only a few items to begin with (e.g., Janssens et al., 1995). Despite these objections, a researcher may feel that he or she can safely delete items from a measurement model without damaging construct validity or undermining theory. The decision about which items to drop, however, can only be made after such items have been properly identified. The procedure described here provides a rigorous and systematic way of identifying them.

The second option is recourse to PFI. Under PFI, the loadings of invariant items are constrained to be equal across groups, while the loadings of non-invariant items are permitted to vary. Byrne et al. (1989) recommend this approach when comparing factor means or structural models under conditions when full invariance cannot be established (e.g., Byrne, 1993; Jackson et al., 1993; Reise et al.,

1993; Riordan & Vandenberg, 1994). Byrne et al. argue that comparison of factor means is still feasible when most of the items are invariant, and that under these conditions, failure to achieve full factorial invariance is trivial from a practical point of view (Marsh & Hocevar, 1985).

Obviously, all non-invariant items must be correctly identified before the researcher can feel assured that "most" items are invariant, or that the effects of non-invariant items are in fact trivial. The procedures described in this paper accomplish this identification task in a rigorous fashion. In the numerical example above, it was shown that using a single, inappropriately chosen referent can prevent identification of a non-invariant item. The extended procedure recommended here is to perform a series of tests using *all* items as referents. In addition, the triangle heuristic not only provides a method for identifying invariant sets, but also a way of estimating in a qualitative fashion *the extent* to which individual items are non-invariant. For example, consider a construct ξ_1 loading on five items, X_1 through X_5. Item X_4 may be noninvariant with respect to one referent; say, X_3. Item X_5, on the other hand, may be noninvariant with respect to two referents, X_3, and X_4. Given this result, and based upon her or his understanding of the construct, a researcher may decide to constrain λ_{41} across groups, while allowing λ_{51} to vary.

Another approach is to treat invariance failure as a source of information concerning meaningful differences between groups. Ellis (1989), for example, used Item Response Theory[6] to examine whether certain measures were non-invariant due to either linguistic or cultural reasons. If invariance failure is due to linguistic problems, then invariance may be achievable by simply rephrasing or retranslating the items. If invariance fails for reasons that cannot be attributed to linguistics, then the researcher may wish to examine the responses as indicators of true cross-cultural differences; in other words, as meaningful data in their own right. As before, accurate identification of non-invariant items is a prerequisite.

In final summary, procedures currently used to identify non-invariant items may lead to incomplete or inaccurate results if single items are used as referents. We propose extensions to the standard procedures used to test for factorial invariance (Byrne et al., 1989). These extensions test all combinations of items and referents for invariance, leading to the identification of sets of invariant items. Invariant sets may be selected for analysis based upon theoretical and substantive considerations. The recommended procedures allow researchers to make more effective use of the various techniques for dealing with invariance, namely, item deletion, partial factorial invariance, and interpretation. Use of these procedures is expected to increase the rigor of cross-cultural research.

Acknowledgments: The authors gratefully acknowledge the assistance provided by the Cross-Cultural Research Methods Group at the Chinese University of Hong Kong, and by two anonymous reviewers whose critical efforts greatly improved the quality of this paper.

Appendix A

Ambiguity Implicit in Type 1 standardization

Type 1 standardization (standardized construct) appears to offer a parsimonious and trouble-free approach to item-level invariance testing, as all the factor loadings can be estimated, and none of them are constrained to equal unity. Unfortunately, Type 1 standardization introduces a source of ambiguity with respect to the global test (Test 2). Consider the four-item, single-construct example shown in Figure 2. An equivalence test using $\phi_{11}^{(1)} = \phi_{11}^{(2)} = 1$ is not a test of the null hypothesis $\lambda_{ij}^{(1)} = \lambda_{ij}^{(2)}$, but rather of $\lambda_{ij}^{(1)}\sqrt{\phi_{11}^{(1)}} = \lambda_{ij}^{(2)}\sqrt{\phi_{11}^{(2)}}$ (Hayduk, 1987). If the test statistic $\Delta\chi^2$ is significant, we have no way of knowing whether this signals the failure of the null hypotheses $\lambda_{ij}^{(1)} = \lambda_{ij}^{(2)}$, or failure of the assumption embodied in the standardization procedure: namely, $\phi_{11}^{(1)} = \phi_{11}^{(2)}$.

In addition, the assumption underlying Type 1 standardization, namely that the variances of samples drawn from different populations are equal, is generally not tenable. The assumption is usually only valid when the different groups are drawn at random from the same population, a situation having little practical relevance to cross-cultural studies.

Appendix B

Technical Rationale for the Factor-Ratio Test

The discussion below uses the four-item model shown in Figure 2. Equations (1) below show the relationships among the variables, which consist of the item variances (σ_{ii}), item covariances (σ_{ij}), factor loadings (λ_{ij}), and the variance of the latent variable (ϕ_{11}). In addition, the equations for item variances include error terms (θ_{ii}).

$$\begin{array}{llll} \sigma_{11} = \lambda_{11}^2\phi_{11} + \theta_{11} & \sigma_{12} = \lambda_{11}\lambda_{21}\phi_{11} & \sigma_{13} = \lambda_{11}\lambda_{31}\phi_{11} & \sigma_{14} = \lambda_{11}\lambda_{41}\phi_{11} \\ & \sigma_{22} = \lambda_{21}^2\phi_{11} + \theta_{22} & \sigma_{23} = \lambda_{21}\lambda_{31}\phi_{11} & \sigma_{24} = \lambda_{21}\lambda_{41}\phi_{11} \\ & & \sigma_{33} = \lambda_{31}^2\phi_{11} + \theta_{33} & \sigma_{34} = \lambda_{31}\lambda_{41}\phi_{11} \\ & & & \sigma_{44} = \lambda_{41}^2\phi_{11} + \theta_{44} \end{array} \quad (1)$$

The only "known" quantities, obtained from the sample data, are the item variances and covariances. All the other variables in the equation set must be estimated, subject to appropriate constraints. Consider the first equation above.

$$\sigma_{11} = \lambda_{11}^2\phi_{11} + \theta_{11} \quad (2)$$

In this equation σ_{11} is a constant, calculated from the experimental data. The term $\lambda_{11}^2\phi_{11}$ is the portion of the variance of X_1 explained by the construct ξ_1. This term represents a particular quantity estimated from the data, and it must remain invariant across whatever different units of measurement may be assigned to the construct. Otherwise, the contribution of the construct to the variance of the manifest variable would not be constant, but would rather depend upon the particular way the construct is being measured (Hayduk, 1987). On the other hand, given any values of σ_{11} and θ_{11}, there are an infinite number of values of λ_{11} and ϕ_{11} that satisfy equation (2). Therefore, λ_{11} and ϕ_{11} can assume different values as required to maintain $\lambda_{11}^2\phi_{11}$ constant across the various units of measurement that may be applied to ξ_1.

If we take X_1 as the referent and set $\lambda_{11} = 1$, then equation (2) can be written

$$\sigma_{11} = (\lambda_{11}')^2\phi_{11}' + \theta_{11}, \text{ where } \lambda_{11}' = 1. \tag{3}$$

Given the requirement that the first term on the right-hand side of the equation be invariant, we can equate (2) and (3) to obtain

$$\begin{aligned} &(\lambda_{11}')^2\phi_{11}' = (1)\phi_{11}' = (\lambda_{11})^2\phi_{11} \\ \text{or} \quad &\phi_{11}' = (\lambda_{11})^2\phi_{11}. \end{aligned} \tag{4}$$

Now consider the second equation in set (1),

$$\sigma_{12} = \lambda_{11}\lambda_{21}\phi_{11}. \tag{5}$$

As before, we wish to specify values of λ_{11}', λ_{21}', and ϕ_{11}' such that $\lambda_{11}' = 1$; that is,

$$\sigma_{12} = \lambda_{11}'\lambda_{21}'\phi_{11}', \text{ where } \lambda_{11}' = 1. \tag{6}$$

Since σ_{12} is a constant, equating (5) and (6) yields

$$\sigma_{12} = (1)\lambda_{21}'\phi_{11}' = \lambda_{11}\lambda_{21}\phi_{11}. \tag{7}$$

But since $\phi_{11}' = (\lambda_{11})^2\phi_{11}$ by equation (4), then

$$\begin{aligned} &\lambda_{21}'(\lambda_{11})^2\phi_{11} = \lambda_{11}\lambda_{21}\phi_{11}; \\ &\lambda_{21}' = \frac{\lambda_{11}\lambda_{21}}{\lambda_{11}^2}; \text{ or} \end{aligned} \tag{8}$$

$$\lambda'_{21} = \frac{\lambda_{21}}{\lambda_{11}}. \tag{9}$$

There are similar expressions for the other factor loadings. Writing out the expressions for all the loadings in the example, including the first (trivial) one, we have

$$\lambda'_{11} = \frac{\lambda_{11}}{\lambda_{11}} = 1; \quad \lambda'_{21} = \frac{\lambda_{21}}{\lambda_{11}}; \quad \lambda'_{31} = \frac{\lambda_{31}}{\lambda_{11}}; \quad \lambda'_{41} = \frac{\lambda_{41}}{\lambda_{11}}. \tag{10}$$

Figure 3 provides a concrete example. Consider the covariance matrix S_1,

$$S_1 = \begin{bmatrix} .3436 & & & \\ .3872 & 1.0144 & & \\ .2904 & .5808 & .7556 & \\ .3872 & .7744 & .5808 & 1.2244 \end{bmatrix}$$

S_1 is used as the input matrix to estimate the parameter values for the two models in Figure 2.

In Model 1, the variance of ξ_1 (i.e., ϕ_{11}) is constrained to equal one. LISREL (Jöreskog & Sörbom, 1993) estimates two of the factor loadings resulting from this constraint as $\lambda_{11} = 0.44$ and $\lambda_{21} = 0.88$. In Model 2, the loading of X_1 on ξ_1 (i.e., λ'_{11}) is constrained to be 1.0. Under this constraint, LISREL estimates the variance of ξ_1 (i.e., ϕ'_{11}) to be 0.1936, while $\lambda'_{21} = 2.0$. Substituting these numbers into equation (4) yields

$$\phi'_{11} = (\lambda_{11})^2\phi_{11} \Rightarrow 0.1936 = (0.44)^2(1) = 0.1936$$

Similarly, using equation (9),

$$\lambda'_{21} = \frac{\lambda_{21}}{\lambda_{11}} \Rightarrow 2.0 = \frac{0.88}{0.44} = 2.0.$$

This example illustrates how the parameters in an identified model can vary. Note that the two models are identical, being estimated from the same covariance matrix; yet the actual values estimated for the parameters depend upon the standardization procedure used.

We now consider item invariance tests of the general two-group model shown in Figure 2. If the model is identified using $\lambda'_{11} = 1$ (i.e., $\lambda^{(1)}_{11} = \lambda^{(2)}_{11} = 1$), then the equalities being tested are (in the new notation)

$$(\lambda^{(1)}_{21})' = (\lambda^{(2)}_{21})' \qquad (\lambda^{(1)}_{31})' = (\lambda^{(2)}_{31})' \qquad (\lambda^{(1)}_{41})' = (\lambda^{(2)}_{41})'$$

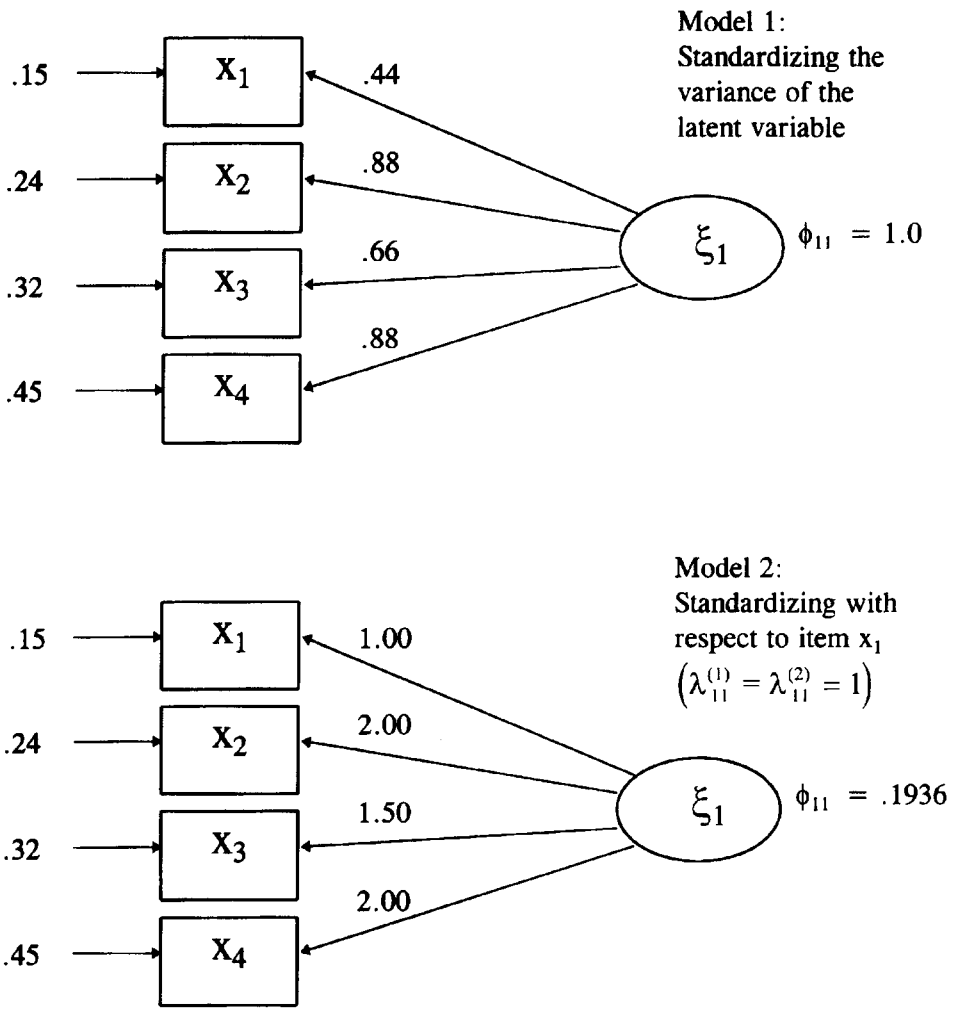

Figure 3. Consequences of Alternative Standardization Procedures

or, alternatively,

$$\frac{\lambda_{21}^{(1)}}{\lambda_{11}^{(1)}} = \frac{\lambda_{21}^{(2)}}{\lambda_{11}^{(2)}} \qquad \frac{\lambda_{31}^{(1)}}{\lambda_{11}^{(1)}} = \frac{\lambda_{31}^{(2)}}{\lambda_{11}^{(2)}} \qquad \frac{\lambda_{41}^{(1)}}{\lambda_{11}^{(1)}} = \frac{\lambda_{41}^{(2)}}{\lambda_{11}^{(2)}}. \tag{11}$$

The choice of referent makes the implicit assumption that $\lambda_{11}^{(1)} = \lambda_{11}^{(2)}$. If $\lambda_{11}^{(1)} \neq \lambda_{11}^{(2)}$, then arbitrarily setting $\lambda_{11}^{(1)} = \lambda_{11}^{(2)} = 1$ makes it impossible to determine whether or not the ratio equalities given above are true. In other words,

if we do not know *a priori* that $\lambda_{11}^{(1)} = \lambda_{11}^{(2)}$, then $\frac{\lambda_{21}^{(1)}}{\lambda_{11}^{(1)}} \neq \frac{\lambda_{21}^{(2)}}{\lambda_{11}^{(2)}}$ does not necessarily imply that $\lambda_{21}^{(1)} \neq \lambda_{21}^{(2)}$. Yet we cannot test $\lambda_{11}^{(1)} = \lambda_{11}^{(2)}$ without choosing some other referent that may, in turn, be non-invariant.

Listing the other relationships implied by equations (11) suggests a solution. The complete set suggests that equivalence can only be established by comparing all ratios of factor loadings. In the example, the following equalities represent all the null hypotheses that must be tested:

$$\begin{aligned}
&\frac{\lambda_{21}^{(1)}}{\lambda_{11}^{(1)}} = \frac{\lambda_{21}^{(2)}}{\lambda_{11}^{(2)}}, \frac{\lambda_{31}^{(1)}}{\lambda_{11}^{(1)}} = \frac{\lambda_{31}^{(2)}}{\lambda_{11}^{(2)}}, \frac{\lambda_{41}^{(1)}}{\lambda_{11}^{(1)}} = \frac{\lambda_{41}^{(2)}}{\lambda_{11}^{(2)}}; \text{ (referent } X_1\text{: } \lambda_{11}^{(1)} = \lambda_{11}^{(2)} = 1) \\
&\frac{\lambda_{31}^{(1)}}{\lambda_{21}^{(1)}} = \frac{\lambda_{31}^{(2)}}{\lambda_{21}^{(2)}}, \frac{\lambda_{41}^{(1)}}{\lambda_{21}^{(1)}} = \frac{\lambda_{41}^{(2)}}{\lambda_{21}^{(2)}}; \text{ (referent } X_2\text{: } \lambda_{21}^{(1)} = \lambda_{21}^{(2)} = 1) \\
&\frac{\lambda_{41}^{(1)}}{\lambda_{31}^{(1)}} = \frac{\lambda_{41}^{(2)}}{\lambda_{31}^{(2)}}; \text{ (referent } X_3\text{: } \lambda_{31}^{(1)} = \lambda_{31}^{(2)} = 1)
\end{aligned} \tag{12}$$

Since equalities of the form $\frac{\lambda_{11}^{(1)}}{\lambda_{11}^{(1)}} = \frac{\lambda_{11}^{(2)}}{\lambda_{11}^{(2)}}$ are trivial and those of the form $\frac{\lambda_{11}^{(1)}}{\lambda_{21}^{(1)}} = \frac{\lambda_{11}^{(2)}}{\lambda_{21}^{(2)}}$ are redundant, the six entries in equation set (12) above represent all the required tests.

Applying the global test for factorial equivalence (Test 2) to the four-item example, it is easy to show that the constraints applicable to the fully constrained model imply the following.

$$\begin{aligned}
&\text{If } \lambda_{11}^{(1)} = \lambda_{11}^{(2)} = 1 \text{ and } \lambda_{21}^{(1)} = \lambda_{21}^{(2)}, \text{ then } \frac{\lambda_{21}^{(1)}}{\lambda_{11}^{(1)}} = \frac{\lambda_{21}^{(2)}}{\lambda_{11}^{(2)}} \\
&\text{If } \lambda_{11}^{(1)} = \lambda_{11}^{(2)} = 1 \text{ and } \lambda_{31}^{(1)} = \lambda_{31}^{(2)}, \text{ then } \frac{\lambda_{31}^{(1)}}{\lambda_{11}^{(1)}} = \frac{\lambda_{31}^{(2)}}{\lambda_{11}^{(2)}} \\
&\ldots(\text{etc.})\ldots; \\
&\text{If } \lambda_{31}^{(1)} = \lambda_{31}^{(2)} \text{ and } \lambda_{41}^{(1)} = \lambda_{41}^{(2)}, \text{ then } \frac{\lambda_{41}^{(1)}}{\lambda_{31}^{(1)}} = \frac{\lambda_{41}^{(2)}}{\lambda_{31}^{(2)}}
\end{aligned} \tag{13}$$

This replicates equation set (12). It is evident that the constraints on the global test do not imply merely a pairwise comparison of factor loadings, but rather a pairwise comparison of *all ratios* of factor loadings. The test does not, for example, determine whether answers obtained from Group 1 and Group 2 subjects are significantly different, but whether or not the *pattern* of answers is different. In other words, it is a test of the equivalence of the covariance structures of item responses across groups.

For the global test, Type 2 standardization is adequate. Equations (13) show that it is permissible to use one referent in each construct, since the constraints imply the entire series of ratio tests. On the other hand, using only one referent is not acceptable when performing item-level tests within a construct. In the example above, it is shown that standardizing with respect to λ_{11} produces a test of *only* the first three equalities in equation set (12). It is possible that the $\Delta\chi^2$ test statistic may not be significant for any of these three tests, leading to the erroneous conclusion that X_2, X_3, and X_4 (and, since we chose it as the referent, X_1) are invariant across groups, even though the global test may have indicated that the overall construct was not invariant . If only X_1 is used, then the tests of the other equalities in equation set (12) are missing, and these omissions may conceal the source of the non-invariance.

Notes

1. This paper utilizes maximum likelihood (ML) estimation, which is commonly used for CFA. Although ML requires manifest variables to possess multivariate normal distribution, it is robust with respect to moderate deviations from this requirement (Bollen, 1989). However, the chi-square estimate is biased under extreme departures from normality. When these conditions exist, researchers should utilize Weighted Least Squares (WLS) estimation. An example of WLS estimation is given by Mullen (1995).
2. Readers are cautioned that when testing factorial invariance, unstandardized coefficients (covariance matrices) ought to be utilized (Bollen, 1989; Singh, 1995). Many researchers automatically rescale their covariance matrices to correlation matrices before performing any type of factor analysis, which is an error when testing factorial invariance across groups. We thank an anonymous reviewer for pointing this out.
3. To make the discussion more comprehensive, χ^2 was used in this paper for comparisons of fit between models. The χ^2 statistic is currently the best available for comparing models because it provides a probability distribution for significance testing. Therefore, researchers can control for experiment-wise error rate for multiple comparisons. However, since χ^2 is sensitive to sample size, researchers may wish to consider changes in other fit indices (e.g., CFI, TLI) in addition to χ^2 as ways to evaluate misfit.
4. Not knowing in advance which constructs may be nonequivalent, we calculated that the experiment could require as many as 23 tests, as follows. One test was required for Test 1, one for Test 2, and one for each construct (3) in Test 3. In addition, each construct was represented by four items, and therefore requires six factor-ratio tests (18). Using a Bonferroni adjustment to control for experiment-wise error at the 0.01 level, we divide 0.01 by 23 and obtain 0.0004. We used a significance level of 0.0001, which is a more commonly used level having this order of magnitude.
5. Finding a theoretical basis for such a choice may be impossible, even for researchers desiring to do so. Many past studies fail to report which items were constrained; even if this information were available, the use of the same constraints with new data sets would have to be justified. (We thank an anonymous reviewer for pointing this out.)
6. For a comparison of multigroup LISREL and IRT as methods for testing factorial invariance, see Reise et al. (1993). A discussion of the topic is outside the scope of this paper.

References

Aiken, L. S., Stein, J. A., & Bentler, P. M. 1994. Structural equation analyses of clinical subpopulation differences and comparative treatment outcomes: Characterizing the daily lives of drug addicts. *Journal of Consulting and Clinical Psychology*, 62: 488–499.

Bandalos, D. L. 1993. Factors influencing cross-validation of confirmatory factor analysis models. *Multivariate Behavioral Research*, 28: 351–374.

Bielby, W. T. 1986. Arbitrary metrics in multiple-indicator models of latent variables. *Sociological Methods and Research*, 15: 3–23.

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

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

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. 1994. Testing for the factorial validity, replication, and invariance of a measurement instrument: A paradigmatic application based on the Maslach Burnout Inventory. *Multivariate Behavioral Research*, 29: 289–311.

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.

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.

Ellis, B. B. 1989. Differential item functioning: Implications for test translations. *Journal of Applied Psychology*, 74: 912–921.

Hayduk, L. A. 1987. *Structural equation modeling with LISREL.* Baltimore: Johns Hopkins University.

ISSP. 1989. *International social science program: Work orientations, 1989* [Computer file]. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributors], 1992.

Jackson, P., Wall, T., Martin, R., & Davids, K. 1993. New measures of job control, cognitive demand, and production responsibility. *Journal of Applied Psychology*, 78: 753–762.

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. 1989. *LISREL 7: A guide to the program and applications.* Chicago, Il: SPSS Inc.

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

MacCallum, R. C., & Tucker, L. R. 1991. Representing sources of error in the common-factor model: Implications for theory and practice. *Psychological Bulletin*, 109: 502–511.

Marsh, H. W. 1993. The multidimensional structure of academic self-concept: Invariance over gender and age. *American Educational Research Journal*, 30 (4): 841–860.

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.

Meredith, W. 1964a. Notes on factorial invariance. *Psychometrika*, 29: 177–185.

Meredith, W. 1964b. Rotation to achieve factorial invariance. *Psychometrika*, 29: 187–206.

Millsap, R. E., & Everson, H. 1991. Confirmatory measurement model comparisons using latent means. *Multivariate Behavioral Research*, 26: 479–497.

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

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.

Poortinga, Y. H. 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.

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

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.

Triandis, H. C. 1994. Cross-cultural industrial and organizational psychology. In H. C. Triandis, M. D. Dunnette & L. M. Hough (Eds.), *Handbook of Industrial and Organisational Psychology*, vol. 4, (2nd ed.). Palo Alto, CA: Consulting Psychologists Press, Inc.

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.

Williams, R., & Thomson, E. 1986. Normalization issues in latent variable modeling. *Sociological Methods and Research*, 15: 24–43.

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.