On assessing stability and change of psychometric properties of multi-item concepts across different situations: A general approach

Taris, T.W.; Bok, I. A.; Meijer, Z.

published in
The Journal of Psychology
1998

DOI (link to publisher)
10.1080/00223989809599169

document version
Publisher's PDF, also known as Version of record

The Journal of Psychology, 1998, *132*(3), 301–316

Assessing Stability and Change of Psychometric Properties of Multi-Item Concepts Across Different Situations: A General Approach

TOON W. TARIS
Department of Social Psychology
Kurt Lewin Institute/Free University Amsterdam

INGE A. BOK
Bok & Taris Psychological Assessment and Research

ZITA Y. MEIJER
Department of Social Psychology
Kurt Lewin Institute/Free University Amsterdam

ABSTRACT. A general procedure for assessing the structural invariance of multi-item concepts across many different situations is described. Specifically, invariance across cultural groups, time, and different media of assessment is addressed. An eight-step procedure that integrates many domain-specific approaches is proposed. Then three applications are presented to demonstrate how the procedure should be used in various substantive domains. Strengths and weaknesses of the procedure are discussed.

ONE RECURRENT THEME in many branches of psychology is how to assess the stability of a particular factor structure across different situations. For example, in longitudinal research, one must know whether the structure of a particular multi-item concept remains unchanged across the waves of a study (Schaubroeck & Green, 1989). Similarly, in a cross-cultural study, a researcher might be interested in examining the stability of a particular factor structure across the cultures involved (Leung & Bond, 1989). Finally, in the area of test development and validation, paper-and-pencil tests are often converted for use on

Zita Meijer is currently affiliated with the Department of Social Psychology, University of West Virginia.

Address correspondence to Toon W. Taris, Free University Amsterdam, Department of Social Psychology, Van der Boechorststraat 1, NL-1081 BT Amsterdam, the Netherlands; fax: +31-20-4448921; e-mail: AW.Taris@psy.vu.nl.

the computer, and the psychometric properties of the transferred test must be retained (Mead & Drasgow, 1993).

In many branches of psychology, procedures have been devised to assess the degree to which factor structures, reliabilities, variances, and means may change across varying situations. Those procedures differ regarding the importance they attach to certain aspects of stability. For example, researchers in developmental and child psychology, fields with a long tradition of longitudinal research, tend to frame issues of stability in terms of stability across *time,* whereas cultural psychologists focus on stability across *cultural groups.* These different foci are both understandable and unfortunate, because the body of procedures developed within one particular domain is often rather similar to, but not quite the same as, the strategies developed in another field. As we show in this article, it is possible to integrate these domain-specific methodologies, leading to a simple (but rigid and systematic) procedure that allows researchers to test all forms of stability and change.

In this article we (a) provide a selective review and integration of earlier approaches to the assessment of stability and change across a variety of situations and (b) propose and illustrate a unifying framework to assess stability and change. Our procedure builds on earlier work in several research areas, including psychometrics (longitudinal), developmental psychology, cultural psychology, and test construction. The applicability of our procedure, however, extends beyond those domains.

In this article, we (a) discuss three types of stability and change of psychometric properties of multi-item concepts; (b) focus more specifically on one type of stability (structural invariance); (c) provide a brief overview of the extended confirmatory factor-analytic model (i.e., with latent mean structures); and (d) propose a general eight-step procedure for examining structural invariance.

Stability of Psychometric Properties of a Concept: Types of Change

In their article on persistence and change in developmental psychology, Mortimer, Finch, and Kumka (1982) addressed three types of fluctuations in stability that may occur across time. First, the level of a particular concept may change—that is, the degree to which the magnitude or quantity of a phenomenon remains unchanged across time. For example, research participants may become more satisfied with their jobs across time. Level change is sometimes also referred to as *alpha* change (Golembiewski, Billingsley, & Yeager, 1976).

Second, Mortimer et al. (1982) discussed the normative stability of a concept—the persistence of individual ranks or differences on an attribute of interest. If one assumes that the nature of the phenomenon of interest has not changed, how does the ordering of individuals with respect to this phenomenon change or persist across time? This type of stability is assessed by computing correlations between measures of the phenomenon in question across time for a group of individuals: High correlation means high stability.

Finally, the factorial structure (structural invariance) of a particular multidimensional concept may change across time. For example, a personality construct may be said to be structurally invariant when it is characterized by the same dimensions and when there is a persistent pattern of relationships among the component attributes over time (Mortimer et al., 1982).

Golembiewski et al. (1976) considered differences regarding the factorial structure of a concept as attributable to either beta or gamma change. Beta change involves a variation in the level of some existential state, complicated by the fact that some intervals of the measurement continuum associated with a constant conceptual domain have been recalibrated. For example, the experiences of new employees during organizational entry have considerable impact on their perceptions, attitudes, needs, and personal reactions, such that newcomers may constitute and reconstitute their interpretations of the work environment during that time (Louis, 1980). An organizational development intervention may give such employees a clearer perception of what the characteristics of the organization actually are. Similarly, a couple in marital therapy may see each other differently after some sessions, even if the situation has factually remained unchanged. In these examples, a change in perspective is involved for the respondents: People may make different estimates of reality, given clearer (or just different) perceptions of what is happening, or they may highlight different aspects of this reality.

Gamma change involves "a redefinition or reconceptualization of some domain, a major change in the perspective or frame of reference within which phenomena are perceived and classified, in what is taken to be relevant in some slice of reality" (Golembiewski et al., 1976, pp. 134–135). For example, it is difficult—if not downright impossible—to compare the order of subjects on a phenomenon that is one-dimensional at one time and multidimensional at another. Thus, whereas beta change involves just a recalibration of the intervals of a scale, gamma change is "big bang" change, implying that comparison across situations is meaningless. As Mortimer et al. (1982) noted, the examination of level and normative stability presumes structural invariance, implying that structural invariance must be examined first before the other forms of stability can be examined, even though the latter are usually of more substantive interest.

Although the discussions of stability and change of Mortimer et al. (1982) and Golembiewski et al. (1976) were geared toward comparison across time, generalization toward other substantive domains is straightforward. For example, if two or more cultures (or groups) are to be compared regarding their scores on a particular multi-item concept (i.e., one would like to examine level differences), the factorial structure of that concept must be the same for all groups under consideration (i.e., structurally invariant). Normative invariance is irrelevant here, because culture is a between-subjects variable.

Similarly, if a psychological test is transferred from one medium to another (usually from paper-and-pencil to computer), one must assess whether the means, standard deviations, and structure of the test are the same across different imple-

mentations, and whether the correlations between measures of the phenomenon in question are high enough to retain the assumption that both implementations tap the same concept (American Psychological Association, 1986). Clearly, this requires an examination of normative, level, and structural invariance.

Assessing Structural Invariance

Substantive domains differ in the importance they attach to the three aforementioned forms of stability. Often, the goal is not so much to find invariance, but to study change. For example, cultural psychologists are interested in the differences among (clusters of) cultures and less so in their similarities. The finding that members of two cultures share the same beliefs, attitudes, or behaviors does not make an interesting study, whereas a difference between cultures often does. However, the quest for publishable data does not usually extend beyond differences regarding means. Therefore, the factorial structure of the concept being compared must be invariant across cultures. Indeed, the assumption (usually implicit) that the relations among the measured items of these concepts are invariant lies at the heart of all research involving comparisons or relations among multi-item concepts.

However, there are many facets involved in establishing structural invariance. One might focus, for example, on the relations among the set of subscales of a multi-item concept, the magnitude of the factor loadings, the reliability of a concept, or the standard deviations of the subscales. Thus, although it is generally acknowledged that structural invariance must come first, it is by no means clear as to *when* it is permissible to say that a particular concept is "structurally invariant" across time, cultural group, or medium. Indeed, it is probably impossible to give a definition of structural invariance that applies to all situations, because different situations tend to highlight different aspects of stability. What is possible, however, is to review critically the requirements that must be met in order to make meaningful comparisons among sets of scores across time, media, or groups, leading to a situation-specific prescription of aspects of structural invariance that must be examined. Of course, one cannot deal with all possible situations. Therefore, our attention is limited to the three situations discussed earlier. These three examples will provide the reader with a good idea of the considerations that one must go through to design an appropriate plan of analysis for one's own situation.

We first consider the confirmatory factor-analytic model. Then we proceed with a proposal for assessing structural, normative, and level stability across a range of situations.

The Confirmatory Factor-Analytic Model

The confirmatory factor-analytic model describes the relations among a set of observed (or manifest) items and a smaller set of unobserved (or latent) variables.

In matrix notation, the relation between an observed variable x_i ($i = 1, \ldots, t$) and a latent variable ξ_j (ksi, $j = 1, \ldots, k$) is

$$x = \Lambda_x \xi + \Theta_\delta, \tag{1}$$

with Λ_x (lambda-x) denoting a $t \times k$ matrix of factor loadings of x on ξ, and Θ_δ (theta-delta) a $t \times 1$ vector of error terms. Thus, the observed score on a particular item x is the result of the weighted score on the underlying latent variable ξ and a particular amount of error. In addition, there is a symmetrical $k \times k$ matrix ϕ (phi) that represents the variance–covariance matrix for the latent variables, and a symmetrical $t \times t$ matrix Θ_δ that gives the variance–covariance matrix for the errors of the observed variables (Jöreskog & Sörbom, 1993). For the current purposes, this model is extended with two extra vectors, namely, a $k \times 1$ vector κ (kappa) that contains the means of the latent variables and a $t \times 1$ vector t_x (tau-x) that presents the means of the observed variables. The relation between observed variable x and latent variable ξ then becomes

$$x = \tau_x + \Lambda_x \xi + \Theta_\delta. \tag{2}$$

The parameters of this model are not identified in single-sample studies, but in multisample studies they can readily be estimated. Thus, if data are available for two or more situations, the means of the latent and the observed variables can be estimated by strategically imposing constraints on parameter values across situations (Jöreskog & Sörbom, 1993). The test statistics are computed across all groups. Note that data on one set of subjects in two or more different situations can usually be arranged in such a way that the parameters of Equation 2 can be estimated. For example, in the case of a longitudinal two-wave study, there is a set of items measured at Time 1 (yielding one sample) and the same set of items measured at Time 2 (yielding a second sample). In effect, a within-subject design is analyzed as if it were a between-subjects design.

Constraining parameters across samples offers the opportunity to test whether imposing such a constraint on a model results in a model that fits the data significantly worse than the unconstrained model, as the constrained model is *nested* within the unconstrained model. The chi-square difference between the constrained and the unconstrained model can be taken as an indication of the deterioration in fit of the constrained model relative to the unconstrained model. If this increase is significant, one would conclude that the samples differ significantly from each other regarding the magnitude of the parameters that were constrained across groups—a procedure that seems simple and viable at first.

This approach, however, is somewhat complicated by the fact that the power of the chi-square test is strongly dependent on the sizes of the samples that are being analyzed (Marsh, Balla, & McDonald, 1988; Saris & Stronkhorst, 1984). In large samples, minor differences among the samples to be compared lead to significant chi-square values when parameters are constrained across groups, whereas in small samples, large differences may remain undetected. This diffi-

culty has led statisticians to develop alternative-fit indices that are less dependent on sample size. One of the best known of those indices is Bentler and Bonett's (1980) nonnormed fit index (NNFI). NNFI is defined as

$$\text{NNFI} = \left(\frac{\chi_n^2}{df_n} - \frac{\chi_t^2}{df_t} \right) / \left(\frac{\chi_n^2}{df_n} - 1 \right), \tag{3}$$

with subscript n referring to a particular null model (usually the model representing independence among the items, that is, the model that assumes that there are no significant relations among the items), and t denoting a particular target model. Values of .90 and over indicate that the target model explains a large enough proportion of the covariance among the items to warrant the conclusion that this model is empirically acceptable (Bentler & Bonett, 1980). Marsh et al. (1988) showed, by means of Monte Carlo simulation, that of 30 well-used fit measures, NNFI is the least dependent on sample size. (Monte Carlo simulations are especially appropriate for studying the behavior of fit indices like NNFI. Although it would be possible to estimate the power of NNFI for large samples numerically, the practical value of such statistical exercises is quite limited because the characteristics of real-life samples tend to deviate from what asymptotic theory assumes. Monte Carlo simulations offer a good opportunity to study the behavior of test statistics in more realistic situations by varying the characteristics of samples.) Thus, it appears that inspection of the NNFI is a useful complement to the usual chi-square test in judging the quality and fit of a particular model.

An Eight-Step Procedure

A natural order in which to examine issues of stability and change of psychometric properties across situations is (1) establish structural invariance, (2) establish level invariance, and (3) establish normative invariance. The order of 2 and 3 is arbitrary, but 1 must always come first, because 2 and 3 are meaningless unless it has been shown that 1 holds. The first step, then, involves a comparison of the factorial structure across various situations.

Structural Invariance

1. One must assess whether the same *number of factors* is present, and whether the *pattern of factor loadings* is the same across situations (an aspect of gamma change). If this assumption does not apply, further analysis is meaningless. Thus, does the basic factorial structure apply to all situations? Has a "simple structure" been reached? Is this structure the same across situations?

2. One must examine whether the *magnitude* of the factor loadings is the same across situations. One would expect that the item loadings are the same across situations. If this assumption is not met, some form of beta change may have occurred (Schmitt, 1982), which implies that a comparison of the factor

scores across situations is not warranted. In practice, this step amounts to constraining the elements of Λ_x to be equal across groups. If the constrained model fits the data considerably worse than an unconstrained model, the magnitude of the factor loadings is not the same across situations.

Note that one may analyze either the correlation or the variance-covariance matrix. In principle, one should analyze covariances rather than correlations. In correlation matrices, the item variances will be equal by definition, which means that it is impossible to find differences between groups regarding the item variances. As such, differences between the groups will be underestimated.

3. Third, are the *covariances* among the scales equal? This question refers to the degree to which respondents see greater integration of the subscales of a concept (gamma change). Even if the number of factors and magnitude/pattern of factor loadings turn out to be the same across situations, it does not follow that the covariances among the subscales are the same. It may well be that, at one point in time, two subscales are more strongly related than at another, implying a shift in the "boundary of meaning" of those concepts. This step involves constraining the off-diagonal elements of the Φ-matrix to be equal across situations.

4. Next, one must test whether the *variances* of the subscales are equal. If no beta change has occurred, the variances of the constructs of interest must be invariant across situations. Changes in factor variances indicate that the respondents perceive more or less difference in the relevant constructs across situations. Sometimes such findings would present a severe conceptual problem (e.g., in the situation where the medium of assessment constitutes the difference between the situations), but there are variations in different situations. For instance, when "culture" constitutes the difference between the groups, there is no reason to expect the variances of a concept to be identical. Sometimes the variance of a particular concept may simply signal a larger disagreement among the respondents in one group than in another, and this finding may have nothing to do with a "recalibration of scale intervals" (Millsap & Hartog, 1988). This step requires that the diagonal elements of the Φ-matrix (representing the variance–covariance matrix for the latent factors) are constrained to be equal across situations.

5. If the variance of a factor is equal across situations, the test of equal *error variances* indicates whether the reliability of the measurement is also equal across situations. Testing the equality of the error variances requires that the diagonal elements of the Θ_δ matrix be constrained to be equal across situations. If the constrained model holds up, the respondents are equally well able to understand and provide answers to the items, regardless of the situations they are in. On the other hand, the reliability may be found to be dependent on the situation (e.g., beta change).

In summary, the first three steps test whether the basic structure of the factor model (including number of factors, and pattern and magnitude of loadings) is invariant across situations. Invariance regarding these three steps is necessary

in order to obtain a meaningful comparison of means across all types of situations. This rule applies to a lesser degree to Steps 4 and 5, during which the equality of factor and error variances across situations is tested. Often these aspects must be stable across situations, but this is not always the case.

If Steps 1–3 did not reveal a significant departure from the assumption of structural invariance, one may proceed with a comparison of means. Steps 6–7 assess whether the means of the latent factors and their observed indicators are equal across situations. Finally, Step 8 addresses the degree of normative stability across situations.

Level Stability

6. As a first step to examine the equality of the latent means across situations, it is convenient to constrain the *means of the items* (i.e., the elements in vector Θ_δ) to be equal across situations. For some comparisons, one would like to see no difference in means (e.g., when two implementations of a test are compared). However, often one neither expects nor desires equality of item means (for instance, in cross-cultural applications or in longitudinal research).

7. Now the equality of the *latent means* (the elements of vector κ) can be tested. Again, in some applications, ideally there would be no difference, whereas in other fields of psychology, such differences would present very interesting findings.

In practice, Steps 6 and 7 cannot be separated. If the item means are different across situations, the means of the latent factors will be also; after all, the latent means are derived from the observed means. Thus, in a way, Steps 6–7 present two faces of the same coin, and one might even consider these two steps as one.

Normative Stability

8. Finally, for some applications it may be desirable to assess the normative stability of a concept. This applies only if a particular concept was measured at least twice for the *same* set of respondents, for example, in longitudinal research (but also when two implementations of a particular test are administered to the same set of respondents; statistically, this is just another repeated measures design).

This step requires a somewhat different set-up of the data. Instead of analyzing two (or more) separate covariance matrices, these matrices must be integrated into one large matrix; only then can within-subject across-situation correlations be computed. For example, if q responses were measured at two occasions, the covariance matrix to be analyzed here is a $2q \times 2q$ matrix; in the

previous steps, we analyzed two separate $q \times q$ matrices, which were treated as statistically independent samples. For this set-up, we must specify the same model as that obtained in Step 5 (with all the appropriate constraints, but now across time). The correlations among the latent factors across situations represent the normative stability of the concepts.

This sequence of eight steps allows one to test systematically the degree to which the psychometric properties of a set of items are stable across a wide range of situations. As indicated earlier, the first five steps relate to structural invariance of a construct. Steps 6 and 7 examine level stability, and Step 8 addresses normative stability.

Three Applications

Example 1: Invariance of Factor Structures Across Cultures

At the heart of much research in cross-cultural psychology lies the assumption that differences among cultures are manifestations of one or more broad underlying dimensions, including masculinity, uncertainty avoidance, and (especially) individualism–collectivism (Hofstede, 1980). Semin, Nandram, Goossens, and Taris (1996) developed a 17-item individualism–collectivism (INDCOL) scale, consisting of three subscales. This scale was used by Meijer and Semin (1996) in their study on cross-cultural differences regarding attributions of success and failure. Their sample consisted of 96 Japanese respondents (M_{age} = 36, SD = 9.7) and 45 Dutch respondents (M_{age} = 33, SD = 13.2). The Japanese were expected to be more collectivistically oriented than the Dutch. Thus, that research involved a comparison of those two cultures via the INDCOL scale (Semin et al., 1996).

For the comparison to be meaningful, the factor structure of the INDCOL scale had to be invariant across groups. However, cultures may differ regarding their location on the INDCOL scale (indeed, the scale was designed to tap such different levels). Similarly, the variance of the dimensions of the scale may well differ from one culture to another; in a culture that is in transition (i.e., from a traditional, collectivist culture toward a more Western-oriented, individualistic culture), one would expect a greater variation in orientation than in a culture that has already made this transition. Thus, the following items must be demonstrated: (a) There should not be any difference regarding the *basic factor structure*: The same number of factors should be present, and the pattern of loadings should be the same. (b) The *magnitude of the loadings* should be identical across cultures, thus warranting that all items have the same weight across cultures. (c) The *covariances among the subscales* should be the same across cultures.

In the Semin et al. (1996) study, we first examined whether the same factorial structure applied to both cultures. A one-factor model without any across-group constraints was strongly rejected, $\chi^2(238, N = 138) = 499.45$, NNFI = .37.

A three-factor model with factor loadings corresponding to the scale developed by Semin et al. (1996), but without across-group constraints, fitted the data considerably better, $\chi^2(232, N = 138) = 407.69$, NNFI = .57. Constraining the loadings to be equal across cultures, however, resulted in a severe deterioration in fit, $\chi^2(246, N = 138) = 439.45$, NNFI = .55; an increase of 31.8 chi-square points with a gain of only 14 *df* was statistically significant at $p < .01$. Thus, the *magnitude* of the loadings was different across cultures (beta change). Finally, when the covariances among the three latent variables were also constrained to be equal across groups, the chi-square increase was not significant, $\chi^2(249, N = 138) = 445.37$, NNFI = .55.

Although the three-factor structure of the latter model did not seem to represent the data very well (NNFI is much below the boundary value of .90), the basic structure of the factor model appeared equal across cultures and, on the subscale level, there were no large differences between the cultures. However, because the absolute magnitude of the loadings differed strongly, one may doubt whether it is really the same concept that is being measured in those cultures. A closer inspection of the data revealed that three items were responsible for the problems encountered here; if those items had been omitted, structural invariance could have been obtained.

Example 2: Equivalence of a Test Across Different Implementations

During the last decade, an increasing number of psychological instruments originally designed as paper-and-pencil tests have been converted for computer use. The same norms must apply to a particular test, regardless of whether a computer or a paper-and-pencil test is used during assessment, meaning that the psychometric properties of a test must be invariant across different implementations. According to the American Psychological Association (1986), two implementations of a test are equivalent if (a) the means, standard deviations, and correlations among the scales are the same across those implementations; (b) the difficulties and reliabilities of the items in one form are the same as the difficulties and reliabilities of the corresponding items in the other form; and (c) the ranking of persons obtained with both forms are the same (American Psychological Association, 1986).

Bok (1990) developed a 21-item paper-and-pencil instrument to tap the quality of work in sheltered workshops. The questionnaire consists of three subscales: (a) Job Content, a 13-item scale; (b) Physical Stresses, a 4-item scale; and (c) Psychic Stresses, a 4-item scale. (Because the last scale was highly correlated to the first, $r = .8$, it is not considered here.) Later, the test developers decided to construct a computerized version of the instrument. Both versions were administered to 38 middle managers, with a repeated measures design in which order of presentation (a within-subject variable) was systematically manipulated. Thus, each respondent completed both versions of the instrument.

To investigate structural invariance, we first examined whether the number of factors and the pattern of factor loadings were the same for both implementations. A comparison of the first three models (see Table 1) showed that the two-factor models fitted the data better than the one-factor model (even though all models fitted the data reasonably well). Because the two-factor model that allowed the two latent factors to be correlated (Model 3) fitted the data significantly better than the model that did not include that correlation (Model 2), we used Model 3 as the starting point for our subsequent analyses.

Successively, we examined whether constraining the loadings to be equal across groups yielded a significant deterioration in fit (Model 4). Our analysis yielded the following: $\chi^2(251, N = 38) = 582.4$, and a slightly lower NNFI of .89. An increase of 43.1 chi-square points with 15 *df* extra was significant at $p < .05$.

TABLE 1
Comparison of Factor Structures Across a Paper-and-Pencil and a Computerized Test, Unweighted Least Squares Estimation

Model number	Model description	Overall χ^2 ($N = 76$)	Overall *df*	NNFI[a]
1	One factor model; all items load on one latent factor, no equality constraints across groups	569.2	238	.89
2	Two-factor model; the items load on one of two uncorrelated (latent) subscales, no equality constraints across groups	547.4	238	.90
3	Model 2 + the two latent subscales are correlated	539.3	236	.90
4	Model 3 + item loadings constrained to be equal	582.4	251	.89
5	Model 4 + covariances, latent variables constrained to be equal across groups	583.6	253	.90
6	Model 5 + variances, latent variables constrained to be equal across groups	586.4	254	.89
7	Model 6 + error variances, items constrained to be equal across groups	603.1	271	.90
8	Model 7 + item means constrained to be equal	617.4	286	[b]
9	Model 8 + means, latent variables constrained to be equal across groups	617.6	288	[b]

Note. $N = 38$.
[a]Bentler and Bonett's (1980) nonnormed fit index. [b]NNFI not computed, because this index pertains only to models without mean structures.

Thus, the two implementations were quite different regarding their item loadings. Closer inspection revealed that this result was attributable to a single item. Because this finding was to be expected on the basis of chance alone, we did not attach much weight to this result.

Then the covariances of the latent variables were constrained to be equal across groups (Model 5). This constraint did not result in a significant deterioration of fit. The same applied when we constrained the variances among the latent variables (Model 6). Thus, the covariance matrix among the latent variables was invariant across groups. Constraining the error variances of the items of the scale to be equal (Model 7) yielded a 16.7-point chi-square increase with 17 *df* extra, which was not significant. All in all, it seemed that there was no strong evidence that both test implementations were not structurally invariant.

To test level invariance, we successively examined whether the item means were the same across the two test implementations (Model 8). Our analysis yielded the following: $\chi^2(286, N = 38) = 617.6$. (Note that this chi-square value can be only loosely compared with the fit of Model 7, because the latter model does not incorporate mean structures.) Constraining the means of the latent variables as well, we found a chi-square increase that was only marginal (Model 8). Thus, the means of the latent variables were invariant across groups as well.

To test for normative invariance, we inspected the intra-test correlations (i.e., the correlation denoting the degree of similarity between the orderings of the persons on the basis of a test, measured with either a computer or paper and pencil). That correlation was .95 or better ($p < .01$) for both scales. Thus, the results obtained with both versions converged to a large degree, regarding the ranking of persons.

In conclusion, this analysis yielded some rather weak evidence that the magnitude of factor loadings was not the same across the two implementations. In all other respects, the hypothesis that both versions tapped the same concept could not be rejected.

Example 3: Invariance of Factor Structures Across Time

The last application refers to the invariance of factor structures across time. The data were collected as part of a longitudinal, two-wave panel study (Fall/Winter 1987/88, and Fall/Winter 1991/92). The respondents were 216 Dutch adults, aged 18–30 years, who were interviewed about their attitudes and behavior regarding several life domains, including work and leisure time. The respondents were attending school at Time 1 and were employed full-time at Time 2. This sample is interesting in that they made the transition from school to work, a transition that is known to affect one's values and attitudes strongly (thus, we assume no *level* stability; cf. Van der Velde, Feij, & Taris, 1995). However, is it also true that the *structure* of these values changes with such a transition? Here we focused on two related attitudinal scales that dealt with the importance the respondent attached to employment. The first scale (three items) tapped the

degree to which a respondent considered work as a duty (WD). Four other items concerned the subjective importance of leisure time (ILP).

Because this sample experienced the transition from school to work, we felt that the relations among the items belonging to a particular scale, as well as the relations among the scales themselves, could have changed. The scales seemed to tap topics whose valence could change with one's employment status; employed people may differ from persons still attending school in their feelings about issues of work as a duty and about leisure time. Therefore, we had reason to apply our eight-step procedure (see Table 2 for the results of our analyses).

First we examined whether the basic structure of the factor model was the same for both time points. A one-factor model was strongly rejected for both time points, $\chi^2(28, N = 216) = 149.54$, NNFI = .49 (Model 1). A two-factor model fitted the data rather well for Time 1, but considerably less so for Time 2 (Model 2). Though NNFI was acceptable (.92), the overall chi-square value was significant ($p < .05$). Thus, it appeared that some form of gamma or beta change had occurred from Time 1 to Time 2. Inspection of the standardized residuals revealed a large residual between two items. After one of those items was omitted, the fit of the unconstrained two-factor model was acceptable, $\chi^2(16, N = 216) = 22.93$, NNFI = .95, Model 3.

When the loadings of the remaining six items were constrained to be equal across time points, the fit of the model did not decrease significantly (Model 4). The same result was found when the covariances between the latent factors were

TABLE 2
Comparison of Factor Structures at Time 1 and Time 2, Unweighted Least Squares Estimation

Model number	Model description	Overall χ^2	Overall *df*	NNFI[a]
1	One factor model; all items load on one latent factor, no equality constraints across time points	149.54	28	.49
2	Two-factor model; the items load on one of two correlated (latent) subscales, no constraints across time points	43.03	26	.92
3	Model 2 + Item 2 omitted	22.93	16	.95
4	Model 3 + loadings constrained to be equal across time points	28.48	20	.95
5	Model 4 + covariances, latent variables constrained across time points	30.87	21	.95

Note. $N = 216$.

constrained (Model 5), $\chi^2(21, N = 216) = 30.87$, NNFI = .95. Thus, there was no reason to assume that a gamma or beta change had occurred; the means of the latent variables could rightfully be compared across time.

Finally, the *normative stability* of the two constructs was assessed. For WD, a correlation of .35 ($p < .01$) was found, and a correlation of .48 ($p < .01$) was found for ILP. Thus, the normative stability of these concepts was not overly high—fitting our expectation concerning the results of experiencing a major life transition.

Conclusion

In the present article, we proposed and illustrated a simple but rigid and systematic eight-step procedure designed to assess whether a particular factor structure is invariant across a variety of situations. Three illustrations were presented, drawn on data from three different strands of psychology. For each application we discussed which steps of the procedure were relevant. To our knowledge, our procedure represents the first attempt to present an integrated procedure to test the equality of factor structures across a variety of frequently recurring situations in many branches of psychology. As such, we believe that our procedure has considerable applicative potential, within psychology as well as in related social and behavioral disciplines.

There are some limitations to our approach. One is that fairly large numbers of respondents are needed, to ensure that the tests are sufficiently powerful to reject the null hypotheses of no difference. However, this limitation is by no means unique to the present procedure. Moreover, our applications demonstrated that even with fairly small samples (fewer than 100 respondents), the tests were sufficiently powerful to detect at least some differences between the groups. One might also turn to test statistics that depend to a lesser degree on sample size than the chi-square test (such as Bentler and Bonett's, 1980, nonnormed fit index). Thus, even if the sample used is rather small, really important differences between factor structures will usually be detected.

One alternative to the procedure proposed here is to use exploratory factor analysis (EFA; e.g., SPSS procedure FACTOR) instead of confirmatory factor analysis (CFA). EFA has the advantage of being considerably easier to use, and many researchers will be much more familiar with EFA than with CFA. However, there are also several drawbacks to EFA. First, it can be used only to compare factor structures, and it does not deal with variances, reliabilities, and covariances among variables. As such, it covers only some of the issues addressed here, and it needs to be complemented by other procedures designed to test the remaining types of stability and change treated here. Second, EFA uses the correlation matrix as input, and, as we noted earlier, this usage implies an a priori standardization of variables, resulting in an underestimation of the differences between situations. Hence, EFA is a poor substitute for CFA as it is used here.

An alternative that may be more appropriate is multilevel confirmatory factor analysis (Hox, 1994; Muthén, 1991). The general idea behind multilevel analysis is that data often have a hierarchical structure. For instance, in our second illustration (invariance of factor structures across cultures), "culture" would be the highest level (we are sampling from cultures; not all cultures were included in the sample), whereas "participant" would be another (lower) level (within each culture, we sampled individuals). In other words, there is a two-level, hierarchical structure here, and this structure can be taken into account via procedures developed by Muthén (1991). Essentially, this approach amounts to separating the between-groups covariation (the covariation between cultures) from the within-group covariation (the covariation within each cultural group) by specifying a LISREL model for both levels. The drawbacks of this procedure are its complexity and its need for fairly large numbers of observations at each level, to ensure sufficient statistical power (as a rule of thumb, one might take 25 observations at each level: thus, 25 cultures would be required; within each culture, we would need 25 participants, yielding a total of at least 625 participants in the application). Many cross-cultural studies use considerably fewer cultures. Thus, it seems that multilevel factor analysis is seldom appropriate.

In summary, we have shown how a simple, eight-step procedure can be used to test many types of stability across a number of different, but frequently occurring, situations in psychological research. The examples presented show how the procedure may be adapted to other situations, and which considerations must play a role. Despite the procedure's limitations—relating mainly to the required size of the sample that is used—it has considerable potential.

REFERENCES

American Psychological Association. (1986). *Guidelines for computer-based tests and interpretation.* Washington, DC: Author.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88,* 588–606.

Bok, I. A. (1990). *Kwaliteit van de arbeid in de sociale werkvoorziening: De ontwikkeling van een instrument ter bepaling van de kenmerken van arbeid* [Quality of work in sheltered workshops: An instrument determining the properties of work]. Amsterdam: Free University, Department of Work and Organisational Psychology.

Golembiewski, R. T., Billingsley, K., & Yeager, S. (1976). Measuring change and persistence in human affairs: Types of change generated by OD designs. *Journal of Applied Behavioral Science, 12,* 133–157.

Hofstede, G. (1980). *Culture's consequences: International differences in work-related values.* Beverly Hills, CA: Sage.

Hox, J. J. (1994). Factor analysis of multilevel data: Gauging the Muthén model. In J. H. L. Oud & R. A. W. van Blokland-Vogelesang (Eds.), *Advances in longitudinal and multivariate analysis in the behavioral sciences* (pp. 141–156). Nijmegen, the Netherlands: ITS.

Jöreskog, K. G., & Sörbom, D. (1993). *LISREL-8* (computer manual). Chicago: Scientific Software.

Leung, K., & Bond, M. H. (1989). On the empirical identification of dimensions for cross-cultural comparisons. *Journal of Cross-Cultural Psychology, 20,* 133–151.

Louis, M. R. (1980). Surprise and sense-making: What newcomers experience in entering unfamiliar organizational settings. *Administrative Science Quarterly, 25,* 226–251.

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

Mead, A. D., & Drasgow, F. J. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114,* 449–458.

Meijer, Z., & Semin, G. R. (1996). *When the self-fulfilling bias does not serve the self: Attributions of success and failure in cultural perspective.* Manuscript submitted for publication.

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

Mortimer, J. T., Finch, M. D., & Kumka, D. (1982). Persistence and change in development: The multidimensional self-concept. In P. B. Baltes & O. G. Brim (Eds.), *Life-span development and behavior* (Vol. 4, pp. 263–313). San Diego: Academic Press.

Muthén, B. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement, 28,* 338–354.

Saris, W. E., & Stronkhorst, H. L. (1984). *Causal modelling in nonexperimental research: An introduction to the LISREL approach.* Amsterdam: Sociometric Research Foundation.

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

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

Semin, G. R., Nandram, S. S., Goossens, A., & Taris, T. W. (1996). *The cultural configuration of emotion: A convergent construct approach.* Manuscript submitted for publication.

Van der Velde, E. G., Feij, J. A., & Taris, T. W. (1995). Stability and change of person characteristics among young adults: The effect of the transition from school to work. *Personality and Individual Differences, 18,* 89–99.

Received November 5, 1996