STRUCTURAL EQUATION MODELING, *12*(3), 471–492

TEACHER'S CORNER

Testing Measurement Invariance of Second-Order Factor Models

Fang Fang Chen, Karen H. Sousa, and Stephen G. West
Arizona State University

We illustrate testing measurement invariance in a second-order factor model using a quality of life dataset ($n = 924$). Measurement invariance was tested across 2 groups at a set of hierarchically structured levels: (a) configural invariance, (b) first-order factor loadings, (c) second-order factor loadings, (d) intercepts of measured variables, (e) intercepts of first-order factors, (f) disturbances of first-order factors, and (g) residual variances of observed variables. Given that measurement invariance at the factor loading and intercept levels was achieved, the latent factor mean difference on the higher order factor between the groups was also estimated. The analyses were performed on the mean and covariance structures within the framework of the confirmatory factor analysis using the LISREL 8.51 program. Implications of second-order factor models and measurement invariance in psychological research were discussed.

Second-order factor models have been used in psychology over a wide variety of domains, including the Big Five personality structure (DeYoung, Peterson, & Higgins, 2002), quality of life (Gotay, Blaine, Haynes, Holup, & Pagano, 2002), self-concept (Marsh, Ellis, & Craven, 2002; Marsh & Hocevar, 1985), psychological well-being (Hills & Argyle, 2002), meaning and satisfaction in life (Harlow & Newcomb, 1990), and HIV stigma (Berger, Ferrans, & Lashley, 2001). Second-order models are most typically applicable in research contexts in which measurement instruments assess several related constructs, each of which is measured by multiple items. The second-order model represents the hypothesis that these seemingly distinct, but related constructs can be accounted for by one or more

Requests for reprints should be sent to Stephen G. West, Department of Psychology, Arizona State University, Tempe, AZ 85287–1104. E-mail: sgwest@asu.edu

common underlying higher order constructs. For example, in a recent application in personality research, evidence suggests that measures of self-esteem, neuroticism, locus of control, and generalized self-efficacy may all be accounted for by a common higher order general factor, and after controlling for the general factor, each trait has little additional ability to predict external criteria (Judge, Erez, Bono, & Thoresen, 2002). In comparison to first-order models with correlated factors, second-order factor models can provide a more parsimonious and interpretable model when researchers hypothesize that higher order factors underlie their data.

Tests of measurement invariance are an important issue if the researcher wishes to make group comparisons (e.g., Byrne & Watkins, 2003; Reise, Widaman, & Pugh, 1993; Van de Vijver & Leung, 1997; Widaman & Reise, 1997). Meaningful comparisons of statistics such as means and regression coefficients can only be made if the measures are comparable across different groups. As one example, the average height of Europeans (measured in meters) and Americans (measured in feet) cannot be directly compared because the two groups are not measured in a common metric.[1] Most applications also assume that the groups are independent and this is the case we consider in this article. Examples of groups on which comparisons are commonly made include gender, age, ethnicity, culture, and experimental versus control groups.

Measurement invariance involves testing the equivalence of measured constructs in two or more independent groups to assure that the same constructs are being assessed in each group. With continuous variables, the most frequently used technique for testing measurement invariance is multiple group confirmatory factor analysis (CFA). The early statistical developments of this technique (e.g., Alwin & Jackson, 1981; Jöreskog, 1971), and the applications that followed, were limited to the comparison of covariance structures. More recent work (Meredith, 1993; Widaman & Reise, 1997) has further developed the technique so that the comparison of mean structures between the groups is also included. This addition is important if investigators intend to go beyond a comparison of the covariance structures in the groups and also compare the mean levels of the constructs, often a question of considerable interest.

Although research examples that examine measurement equivalence across groups considering both the covariance and mean structures are becoming increasingly common (e.g., Kim, Brody, & Murry, 2003; Leone, Perugini, Bagozzi, Pierro, & Mannetti, 2001; Li, Harmer, Acock, Vongjaturapat, & Boonverabut, 1997; Reise, Smith, & Furr, 2001), few examples exist that test measurement invariance in second-order models. The few examples that do exist (e.g., Byrne, 1995; Byrne & Campbell, 1999; Marsh & Hocevar, 1985) do not go beyond exami-

[1]In this case 1 m is known to be approximately 3.28 ft. However, in general, no similar conversion rule exists with which to equate the metrics of two factors for which measurement invariance does not hold.

nation of the covariance structure. In this article we illustrate how to test measurement invariance in multiple group models using the full mean and covariance structures, based on techniques developed by Meredith (1993) and Widaman and Reise (1997). Measurement invariance across groups at the configural, factor loading, intercept, residual variance, and disturbance levels was tested using a second-order factor model. Given an adequate level of measurement invariance, latent factor mean differences between the groups were further tested. Our analyses were performed within the framework of CFA by using the LISREL 8.51 program (Jöreskog & Sörbom, 1999). Other structural equation modeling programs such as Amos, EQS, M*plus,* and MX can also be used to test measurement invariance.

SECOND-ORDER FACTOR MODELS: APPLICABILITY AND ADVANTAGES

Second-order models are potentially applicable when (a) the lower order factors are substantially correlated with each other, and (b) there is a higher order factor that is hypothesized to account for the relations among the lower order factors. For example, to test whether there is a general intelligence factor that underlies a wide range of specific intelligence-related abilities (Spearman, 1927), we can hypothesize that the specific abilities (which are each assessed by multiple items) are lower order factors, and the general intelligence is a higher order factor, which accounts for the commonality among the specific abilities. Statistical tests of the fit of a hypothesized second-order factor normally require that four or more first-order factors are included in the dataset.[2]

A second-order factor model has several potential advantages over a first-order factor model. First, the second-order model can test whether the hypothesized higher order factor actually accounts for the pattern of relations between the first-order factors. Second, a second-order model puts a structure on the pattern of covariance between the first-order factors, explaining the covariance in a more parsimonious way with fewer parameters (Gustafsson & Balke, 1993; Rindskopf & Rose, 1988). Third, a second-order model separates variance due to specific factors from measurement error, leading to a theoretically error-free estimate of the specific factors. The unique variance of each first-order factor that is not shared by the common second-order factor represents the specific factors. These specific factors are represented by the disturbance of each first-order factor. Finally, second-order factor models can provide useful simplification of the interpretation of

[2]With a single second-order factor, three first-order factors result in a model that is just identified and four first-order factors result in a model in which the second-order portion of the model is overidentified. Only when the second-order portion of the model is overidentified can it be properly tested for fit.

complex measurement structures such as multitrait–multimethod models (Eid, Lischetzke, Nussbeck, & Trierweiler, 2003) and latent state–trait models (Steyer, Ferring, & Schmitt, 1992).

MEASUREMENT INVARIANCE

Tests of measurement invariance examine whether the same construct has been measured across different groups. Measurement invariance can be tested at different levels and Meredith (1993) and Widaman and Reise (1997) described procedures for testing a hierarchical series of models to establish measurement invariance in first-order factor models. They developed a specific hierarchical structure of the tests to maximize the interpretability of the results at each step of the hierarchy. Following a review of their work, we discuss additional tests that must be included for a complete test of a second-order model.

First-Order Models

The most basic level of measurement invariance is configural invariance (Horn, McArdle, & Mason, 1983). The central requirement is that the same item must be an indicator of the same latent factor in each group; however, the factor loadings can differ across groups. When this level of invariance is achieved, similar, but not identical, latent variables are present in the groups (Widaman & Reise, 1997).

The second level of invariance is factor loading invariance. Factor loadings represent the strength of the linear relation between each factor and its associated items (Bollen, 1989; Jöreskog & Sörbom, 1999). When the loading of each item on the underlying factor is equal in two (or more) groups, the unit of the measurement of the underlying factor is identical. Of importance, this level of invariance does not require that the scales of the factors have a common origin. When this level of invariance is met, relations between the factor and other external variables can be compared across groups, because one unit of change in one group would be equal to one unit of change in another. However, the factor means of the scale still cannot be compared across groups, as the origin of the scale may differ. A commonly used illustration is the measurement of temperature using the Kelvin scale in Group A and the Celsius scale in Group B (Van de Vijver & Leung, 1997). The measurement unit is identical in both groups but the origins of the scales are not. By subtracting 273 from the temperatures in Celsius in Group B, temperature in Group B will be converted into degrees Kelvin, which is used in Group A.

The third level of invariance is intercept invariance. Intercepts represent the origin of the scale. In testing this form of invariance, intercepts of the measured variables are constrained to be equal across groups, in addition to factor loadings of the latent variables. This level of invariance is required for comparing latent mean dif-

ferences across groups (Widaman & Reise, 1997). When this level of invariance is achieved, it means that scores from different groups have the same unit of measurement (factor loading) as well as the same origin (intercept), and thus the factor means can be compared across groups. Otherwise, it cannot be determined whether any difference between groups on factor means is a true group difference or a measurement artifact.

The fourth form of invariance is residual invariance. In testing this form of invariance, the residual (uniqueness or measurement error[3]) associated with each measured variable is constrained to be equal across groups, in addition to the loadings of the latent variables and the intercepts of the measured variables. When this level of invariance holds, all group differences on the items are due only to group differences on the common factors. Residual invariance, however, can be difficult to achieve for a variety of reasons (see Widaman & Reise, 1997).

Within the hierarchy of tests advocated by Widaman and Reise (1997), measurement invariance in first-order models can be tested at more advanced levels, such as variance and covariance invariance. However, these more advanced levels of invariance represent very stringent ideal standards that are extremely difficult to fulfill. Thus, configural invariance, factor loading, intercept, and residual invariance are the most commonly tested forms of invariance for first-order factor models. In some applications, arguments can be made for placing a greater emphasis on testing of factor covariances and factor variances for invariance. Marsh (1994), McArdle and Nesselroade (1994), Meredith (1993), and Widaman and Reise (1997) discussed procedures and issues associated with such tests. However, factor variances are not generally expected to be equal in different populations.

Second-Order Models

There are important additional aspects of testing measurement invariance of second-order models. First, factor loading invariance must be tested for both the first-order and second-order factors. Second, intercept invariance must be tested for both the measured variables and first-order factors. The first-order factor means are a function of the intercepts of the measured variables and the first-order factor loadings and means; the second-order factor mean is a function of the intercepts of the first-order factors, and second-order factor loadings and means. Finally, in addition to testing the invariance of the residual variance of the observed variables, the invariance of the disturbances (specific factors) of the first-order factors must also be tested. When this level of invariance is achieved, it means the disturbances (specific factors, e.g., spatial ability in the case of intelligence) of the lower order factors are equivalent across the groups.

---

[3]The residual variance or uniqueness of the measured item is composed of both item-specific unique variance and random measurement error.

Latent Mean Difference Test

Multigroup CFA may also be used to test whether the latent factor means differ across the groups. In a typical covariance structure model (Hoyle, 1991), the covariance matrix is computed from deviation scores so that the means of all measured variables will be zero. As a result, the means of all latent constructs are assumed to be zero. To test the latent construct mean differences, a combined mean and covariance structure model must be used (Bentler, 1989; Bollen, 1989; Sörbom, 1978). To estimate the difference between the factor means, one group is usually chosen as a reference or baseline group and its latent means are set to zero. The latent means of the other group, which actually represent the difference between the factor means in the two groups, are estimated. The significance test (Wald or *z* test) for the latent means of the second group provides a test for significance of the difference between the means of the two groups on the latent construct[4] (Aiken, Stein, & Bentler, 1994). In the following section, we illustrate the test of measurement invariance and mean difference tests across two groups for a second-order factor model of quality of life measurement in HIV patients.

METHOD

The factor structure of a 17-item health-related quality of life measurement from the AIDS Time-Oriented Health Outcome Study (ATHOS) was examined. ATHOS is a longitudinal observational database of people with HIV-associated illness cared for by community-based providers in the greater San Francisco, Los Angeles, and San Diego areas. The measurement of quality of life is composed of four subscales: cognition, vitality, mental health, and health worry. Items comprising these subscales were derived from general health status scales (Stewart & Ware, 1992). The full set of items is given in the caption for Figure 1. Items that were reverse-coded are denoted by (R). Items were answered on a 5-point scale ranging from 1 (*all of the time*) to 5 (*never*) so that high scores on the scale represent a high quality of life. We used the data from each participant's initial measurement following his or her enrollment in the study.

The 5-item cognition subscale assesses day-to-day problems in cognitive functioning of which the patient would be aware. The 4-item vitality subscale assesses the energy–fatigue continuum (Stewart, Hays, & Ware, 1992). The 5-item mental health subscale is designed specifically to measure psychological distress and

[4]The Wald test is only approximate. Another, less convenient approach is to estimate a second model in which the latent means of the second-order factors are constrained to be equal. The chi-square difference test comparing the fit of the constrained and unconstrained models provides a more accurate test of the latent means.

well-being. The 3-item health worry subscale measures the extent to which health problems cause people to worry or be greatly concerned about their health (Stewart et al., 1992).

For ease of presentation, we limited our sample to participants for whom complete data were available on all items. The sample size was 924; 84% were White and 95% were male. For the purposes here, participants were classified into two independent groups based on their working status. Working status represents an important categorical distinction that we expected to be related to well-being, particularly in this sample of HIV/AIDS patients. People classified in the working group ($n = 476$) reported working full time or working part time. All other responses (unemployed, laid off, or looking for work; disabled and no longer working; retired; keeping house; or other) were coded as the nonworking group ($n = 448$).

It was hypothesized that there was a second-order factor structure for the quality of life instrument, with cognition, vitality, mental health, and health worry as the lower order factors, and global quality of life as the higher order factor (Ware, Davies-Avery, & Brook, 1980). This structure is depicted in Figure 1.

ANALYSIS OF THE DATA

Tests for the factor structures of the overall quality of life measurement, for its invariance across working and nonworking groups, and for the latent mean differences were based on the analysis of mean and covariance structures, within the framework of CFA. Analyses were conducted using the LISREL 8.51 program (Jöreskog & Sörbom, 1999) and maximum likelihood estimation procedures. Maximum likelihood procedures were used because initial examination of the data did not show evidence of excessive nonnormality (skewness: $Mdn = -.32$, range = –.71 to .07; excess kurtosis: $Mdn = -.29$, range = –.44 to .08).

The analysis followed two major stages. First, measurement invariance was hierarchically tested at each of the levels: configural invariance, invariance of the factor loadings, invariance of the intercepts, and invariance of residual variances (Meredith, 1993; Widaman & Reise, 1997). Following the logic of Widaman and Reise (1997), the disturbances of the first-order factors were also tested, tests that are unique to the second-order models. Second, given that both factor loadings and intercepts were invariant, the mean differences on the higher order latent factor were tested.

Specification of the Hypothesized Model

The hypothesized second-order factor model presented in Figure 1 was specified in the following way: (a) each item would have a nonzero loading on the first-order factor (cognition, vitality, mental health, disease worry) that it was designed to

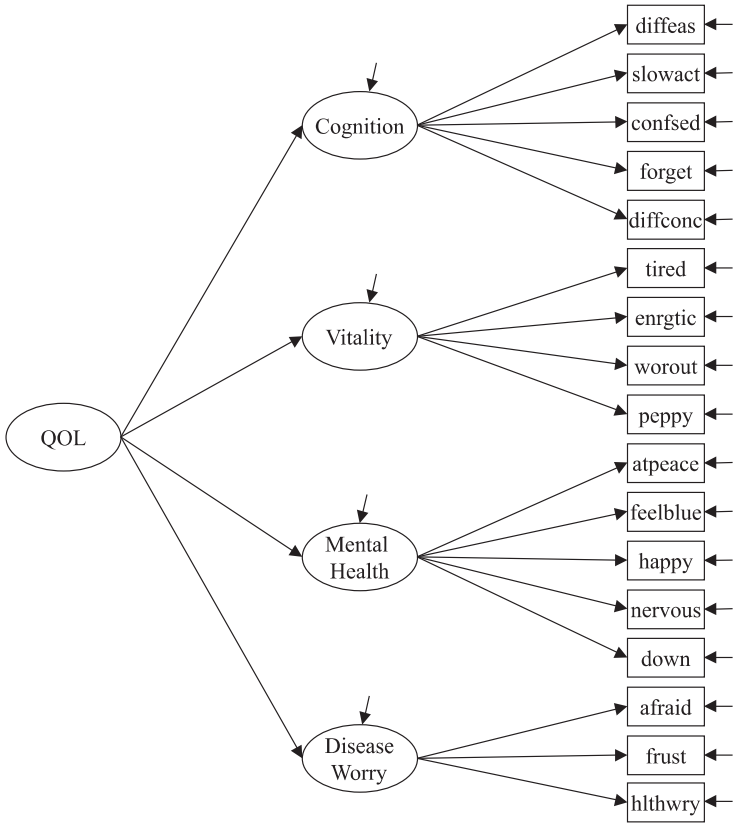

FIGURE 1 Measurement model for the second-order factor model of quality of life. *Note*. Items with acronym in parentheses. Items that were reverse coded are denoted with (R). (a) Cognition subscale: "Have difficulty reasoning and solving problems?" (diffeas); "React slowly to things that were said or done?" (sloact); "Become confused and start several actions at a time?" (confused); "Forget where you put things or appointments?" (forget); "Have difficulty concentrating?" (diffconc). (b) Vitality subscale: "Feel tired?" (tired); "Have enough energy to do the things you want?" (R) (enrgtic); "Feel worn out?" (wornout); "Feel full of pep?" (R) (peppy). (c) Mental health subscale: "Feel calm and peaceful?"(R) (atpeace)? "Feel downhearted and blue?" (feelblue); "Feel very happy"(R) (happy); "Feel very nervous?" (nervous); "Feel so down in the dumps nothing could cheer you up? (down). (d) Disease worry subscale: "Were you afraid because of your health?" (afraid); "Were you frustrated about your health?" (frust); "Was your health a worry in your life?" (healthwry). Second order factor: Quality of life (QOL)—High scores indicate high quality of life.

measure and a zero loading on each of the other first-order factors; (b) error terms associated with each item would be uncorrelated; and (c) all covariance between each pair of the first-order factors would be explained by a higher order factor—which we term *global quality of life.* In LISREL, all first-order factor loadings were defined as LYs, and all second-order factor loadings were defined as GAs. The annotated LISREL computer script is available at http://psych.asu.edu/people/faculty/swest.html.

Identification of the model is required for estimation in CFA. A model is identified if there is a unique numerical solution for each of the parameters (Ullman, 2001). There are two approaches that are typically used to identify the scale of measurement models: One is to fix one of the factor loadings (marker variable) to a value of 1 for each factor, and the other way is to fix the variance of each factor to 1, which standardizes the factor loadings within each group. It is possible for the factor loadings to be invariant even though the true factor variances differ in the two populations (see Cudeck, 1989; Meredith, 1993). We used the marker variable strategy for ease of interpretation. The central issue facing this strategy is which indicator should be chosen as the marker variable. In cases in which an indicator variable is available that is measured in a clear metric (e.g., income in dollars, weight in kg), this variable should normally be chosen as the marker. Or, if an indicator is known from prior research to be invariant across populations, it is also a good choice for the marker variable. In this case we had no basis for choosing a marker so we arbitrarily designated the first indicator of each construct to be the marker. We also explored the use of alternative marker variables and found no material differences across solutions.[5] We used the same strategy with the second-order factor loadings. The fit statistics for each model are presented in Table 1, and the unstandardized factor loadings corresponding to the baseline configural model (Model 1) are presented in Figure 2, Panel A for the working group and Panel B for the nonworking group.

Testing Invariance Across Groups: Comparative Model Testing

In comparing the fit of hypothesized models, chi-square tests and goodness-of-fit indexes are used. The chi-square test assesses the magnitude of the discrepancy between the sample and fitted covariance matrices. A significant test result indicates a poor fit. However, moderate discrepancies from normality in the data also lead the chi-square test to reject the model (West, Finch, & Curran, 1995). Also, when the sample size is large, a small discrepancy from the model that may be of no practical or theoretical interest can lead the chi-square test to reject the model. Consequently, we

[5]Reise et al. (1993) presented an extensive discussion of possible alternative strategies for identifying multiple group models.

TABLE 1
Summary of Fit Statistics for Testing Measurement Invariance of Second-Order Factor Model of Quality of Life

Model[a]	χ^2	*df*	*RMSEA*	*SRMR*[b] *Working*	*SRMR*[b] *Non-working*	*CFI*	*Model Comparison*	$\Delta\chi^2$	Δdf
Model 1 configural invariance	742.60	230	.07	.04	.06	.95	—	—	—
Model 2 first-order factor loadings invariant	757.16	243	.07	.05	.07	.95	2 vs. 1	14.56	13
Model 3 first- and second-order factor loadings invariant	758.31	246	.07	.05	.06	.95	3 vs. 2	1.15	3
Model 4 first- and second-order factor loadings and intercepts of measured variables invariant	838.45	259	.07	.05	.06	.94	4 vs. 3	80.14*	13
Model 5 first- and second-order factor loadings, and intercepts of measured variables and first-order factors invariant	891.83	262	.07	.06	.07	.93	5 vs. 4	53.38*	3
Model 6 first- and second-order factor loadings, intercepts, and disturbances of first-order factors invariant	930.60	266	.08	.07	.07	.93	6 vs. 5	38.77*	4
Model 7 first- and second-order factor loadings, intercepts, disturbances of first-order factors, and residual variances of measured variables invariant	1106.61	283	.08	.07	.08	.91	7 vs. 6	176.01*	17

Note. RMSEA = root mean squared error of approximation; SRMR = standardized root mean square residual; CFI = Comparative fit index.

[a]$N = 924$; 476 vs. 448. LISREL reports a separate SRMR for each group. The SMSRs for each group can be weighted by the group's sample size to form an overall weighted mean SRMR.

*$p < .001$

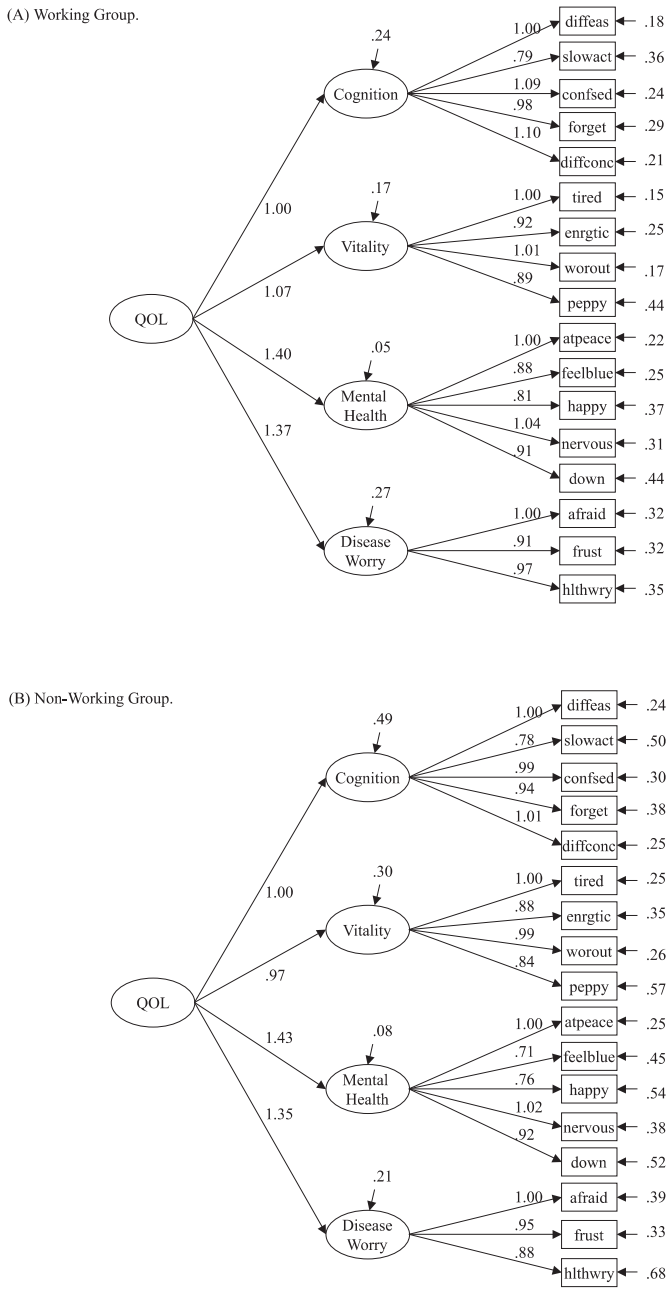

FIGURE 2 Results of the second-order factor model: Unstandardized solution. (A) Working group; (B) nonworking group. *Note.* For the working group, the R^2 for the first-order factors were cognition (.48), vitality (.60), mental health (.90), and disease worry (.61). For the nonworking group, the R^2 for the first-order factors were cognition (.37), vitality (.48), mental health (.88), and disease worry (.72).

also report three fit indexes that showed good performance in a simulation study by Hu and Bentler (1998). The root mean squared error of approximation (RMSEA; Steiger, 1990) is a measure of the estimated discrepancy between the population and model implied population covariance matrices per degree of freedom. Browne and Cudeck (1993) suggested that values of the RMSEA of .05 or less indicate a close fit, and .08 or less indicate adequate fit. The standardized root mean square residual (SRMR; Hu & Bentler, 1998) is a measure of the average of the standardized fitted residuals. It ranges from .00 to 1.00, and a value of less than .08 indicates a good fit. The Comparative fit index (CFI; Bentler, 1990) ranges from 0 (poor fit) to 1.00 (perfect fit) and is derived from a comparison of a restricted model (one in which structure is imposed on the data) with a null model (one in which all pairs of observed variables are assumed to be mutually uncorrelated). The CFI provides a measure of complete covariation in the data. Originally, a value .90 or greater was suggested as evidence of adequate fit, but Hu and Bentler (1999) more recently suggested the use of .95 as a criterion for adequate fit.

In the context of testing measurement invariance, a series of hierarchically nested models are tested. Each pair of models in the sequence is nested because a set of parameters are constrained to be equal across groups in the more restricted but not in the less restricted model. For example, in the configural invariance model, no constraints are placed on the values of the hypothesized factor loadings across groups, whereas the factor loading invariance model constrains the factor loadings to be equal in each group. To compare the fit for two nested models, the chi-square difference (likelihood ratio) test is used (Bentler & Bonett, 1980). If the chi-square difference test is significant, it suggests that the constraints on the more restricted model may be too strict. Otherwise stated, the more restricted model fails the test of measurement invariance across groups and the results of the less restricted model should be accepted. However, once again, the performance of the chi-square difference test is also affected by nonnormality and large sample size so that goodness-of-fit indexes are typically also used to assess model fit. Assessment of fit is an active area of research and no definitive, widely accepted guidelines have yet been established in the context of testing measurement invariance. At present, the best available guidelines are probably those proposed by Cheung and Rensvold (2002). Based on the results of the only large-scale simulation study to date, they concluded that a difference of larger than .01 in the CFI would indicate a meaningful change in model fit for testing measurement invariance.[6] We followed

[6]Cheung and Rensvold (2002) specifically recommend Gamma hat, McDonald's delta noncentrality index, and the CFI to evaluate measurement invariance. Only the CFI is currently available in the output of most structural equation modeling programs (e.g., LISREL, M*plus*). Cheung and Rensvold did not include the SRMR in their evaluation and their specific simulation procedure did not permit a full evaluation of the RMSEA. A disadvantage of the CFI is that it is relatively insensitive to mean structure that is particularly important in testing the invariance of the intercepts (see discussion later in the article).

their general recommendation in this article and used both the chi-square difference test and change in the value of the CFI to evaluate model fit. Following Hu and Bentler (1998, 1999), we also report for readers' interest the values of the RMSEA and SRMR in Table 1. The conclusions that would be reached based on these fit indexes would not differ from those reached based on the CFI.

To test whether the second-order factor structure is statistically equivalent across the two groups, a hierarchical series of nested models were tested, following the general procedures suggested by Widaman and Reise (1997).

Configural invariance (Model 1). In testing for this form of invariance, an unrestricted baseline model was specified in which each group had the same structure. That is, the pattern of fixed and free factor loadings for the first- and second-order factor loadings was constrained to be the same across groups, but different estimates were allowed for the corresponding parameters in the different groups. As can be seen from Table 1 the χ^2 statistic was 742.60 ($df = 230$), $p < .001$. RMSEA was .07, SRMR was .05, and CFI was .95. These results indicated an adequate fit of the model to the data.

Invariance of first-order factor loadings (Model 2). In testing for this level of factorial invariance, all of the first-order factor loadings were constrained to be equal across groups. This level of invariance was nested within Model 1. As can be seen from Table 1, the chi-square difference test was not significant, $\chi^2_\Delta(\Delta df = 13)$, = 14.56, *ns*. These results indicated that the first-order factor loadings were invariant across the working and nonworking groups.

Invariance of second-order factor loadings (Model 3). In testing for this level of invariance, all first- and second-order factor loadings were constrained to be equal across groups. This form of invariance is nested within Model 2. The chi-square difference test was not significant, $\chi^2_\Delta(\Delta df = 3) = 1.15$, *ns.* These results indicated that the second-order factor loadings were invariant across the working and nonworking groups.

Invariance of intercepts of measured variables (Model 4). Models 4 and 5 impose additional constraints to determine whether two different sets of intercepts are invariant. In Model 4 the focus is on the measured variables. In addition to the constraints already imposed on the first- and second-order factor loadings in Model 3, the intercepts of the measured variables were constrained to be equal across groups. This condition is required to detect potential differences in the intercepts of the measured variables between groups when only the first-order factors are involved. The chi-square difference test between Model 4 and Model 3 was significant, $\chi^2_\Delta(\Delta df = 13) = 80.14$, $p < .001$. Given that the test was based on a large sample size for psychological research ($n = 924$), and there was no substantial dif-

ference in CFI (.95 vs. .94), we concluded that there was no appreciable difference between the working and nonworking groups on the intercepts of the measured variables. (We consider the implications of using the chi-square difference test rather than the CFI as the criterion later.)

Invariance of intercepts of first-order latent factors (Model 5). In a second-order factor model, the intercepts of the first-order latent factors must also be invariant across groups in addition to intercept invariance of measured variables to compare the second-order factor means across groups. In testing for this level of invariance, first- and second-order factor loadings and the intercepts of the measured variables and first-order latent factors were constrained to be equal across groups. The chi-square difference test between Models 4 and 5 was significant, $\chi^2_\Delta(\Delta df = 3) = 53.38$, $p < .001$. Once again, given that there was no substantial difference in CFI (.94 vs. .93), we concluded that there was no appreciable difference in the intercepts of the first-order factors across the two groups.

Invariance of disturbances of first-order factors (Model 6). In testing for this level of factorial invariance, all first- and second-order factor loadings, the intercepts of the measured variables and the first-order factors, and disturbances of the first-order factors were constrained equal across groups.[7] This model is nested within Model 5 and the chi-square difference test between the two models was significant, $\chi^2_\Delta(\Delta df = 4) = 38.77$, $p < .001$. Once again, there was no substantial difference in CFI (.93 vs. .93), so we concluded that there was no appreciable difference in the disturbances, which are the unique variances that are not shared by the common higher order factor between the working and nonworking groups.

Invariance of residual variance of observed variables (Model 7). In testing for this level of factorial invariance, all first- and second-order factor loadings, the intercepts of the measured variables and the first-order factors, disturbances of the first-order factors, and residual variances of the measured variables were constrained equal across groups. The chi-square difference test between Model 7 and Model 6 was significant, $\chi^2_\Delta(\Delta df = 17) = 176.01$, $p < .001$, indicating a significant difference between the working and nonworking group on the residual variance of the observed variables. For this comparison, the CFI also indicated that a substantial change in fit had occurred (.93 vs. .91). This result indicates that the residual variances of the measured variables were not invariant across the two groups. Model 6 in which the first- and second-order factor loadings, the intercepts of the

[7]Models 6 and 7 can theoretically be tested in either order. We tested Model 6 first because of the substantially greater theoretical interest in the invariance of the disturbances of the first-order factors (the specific factors) than the uniqueness of the measured variables.

measured variables and the first-order factors, and the disturbances of the first-order factors were constrained to be equal represented the highest level of invariance that could be achieved with these data.

Implications of choice of criterion. The conclusions of the analysis depend on the criterion selected. If we rely on the results of the chi-square difference test, then invariance of the first-order and second-order factor loadings is achieved. Indeed, based on this criterion, we would not conduct any further tests once we had determined that there was a significant difference between Model 4 and Model 3. Model 3 in which the loadings of the first- and second-order factors were constrained to be equal would represent the highest level of invariance that could be achieved. On the other hand, if we rely on the criterion of the change in the CFI suggested by Cheung and Rensvold (2002), then Model 6 represents the highest level of invariance that could be achieved. Of importance, the comparison of the higher order factor mean across groups[8] that we report below requires that invariance be achieved for the first- and second-order factor loadings and for the intercepts of the measured variables and first-order factors (through Model 5). Using the Cheung–Rensvold guidelines, this requirement was surpassed, and we report this comparison later. We will return to the issue of the choice of a criterion for measurement invariance in the discussion.

Stage 2: Test of the Group Difference on the Second-Order Factor Mean

To obtain an estimate of the difference between the higher order factor means in the two groups, the working group was chosen as a reference or baseline group and its second-order latent mean was set to zero. The latent mean of the nonworking group was estimated; this value reflects the difference between the factor means of the two groups. The significance test (Wald z test) for the latent means of the nonworking group is the test for significance of the difference between the means of the two groups on the latent construct (Aiken et al., 1994; Sörbom, 1978).

Invariance of first- and second-order factor loadings, and intercepts of the measured variables and first-order factors was imposed on the working and nonworking groups. There was a significant mean difference between the two groups on the higher order factor (–.49, $z = -11.29$, $p < .0001$), indicating that the nonworking group had a lower score on the global quality of life factor than the working group.

[8]The first-order factor means are conditional on the higher order factor mean(s) in a hierarchical model, and thus cannot be directly compared.

Summary

The second-order factor model of quality of life fit the working and nonworking groups adequately. The tests of measurement invariance using the Cheung–Rensvold criteria indicated that all first- and second-order factor loadings, intercepts of the measured variables and first-order factors, and disturbances of the first-order factors were equivalent. Given an adequate level of measurement invariance, the difference between the group means on the higher order factor was tested, and it was found that the working group had a higher level of overall quality of life than the nonworking group.

DISCUSSION

This article illustrated the strategy of testing measurement invariance in a second-order model of quality of life. The second-order factor model hypothesized that the responses to the measurement of quality of life could be explained by four first-order factors (cognition, vitality, mental health, and disease worry). Moreover, there was one second-order factor (global quality of life) that underlies the four first-order factors. Compared to a correlated four-factor model, the second-order factor model is more parsimonious and provides theoretically error-free estimates of both the general factor and each specific factor. This latter advantage is particularly important when researchers are interested in understanding whether the specific factors can predict external criteria over and above the general factor.

Measurement invariance provides strong evidence that the same construct has been measured across different groups, and it is an important issue in comparing results across groups. Our illustration tested measurement invariance across groups for the second-order factor model at seven different hierarchical levels that are tested in sequence: configural invariance, factor loadings of the first-order factors, factor loadings of the second-order factors, intercepts of observed variables, intercepts of the first-order factors, disturbances of the first-order factors, and residual variances of the observed variables.

Compared to first-order factor models, test of measurement invariance for second-order models are more complex. To ensure that the unit of the scale is the same, both the first-order factor loadings and second-order factor loadings must be invariant. This level of invariance is required to compare the relation between the constructs (e.g., unstandardized regression coefficients). Similarly, to additionally demonstrate that the origin of the scale is the same, both the intercepts of the observed variables and first-order factors must be shown to be invariant. This level of invariance permits the researcher to test the differences in factor means across the groups. Otherwise, the factor mean differences are potentially confounded with

possible differences in the participants' origins of the scales in the different groups. The invariance of disturbances of the first-order factors as well as the invariance of the uniqueness of the measured variables may also be tested. The former test may be of particular interest in second-order models as it ensures that the unique variance of each lower order factor that is not shared by the common higher order factor is the same. The latter test assures that the set of common factors are entirely responsible for any observed differences on the means of the measured variables. These more advanced levels of invariance are difficult to achieve and are not required for testing differences in relationships or mean differences between the groups. Indeed, the invariance of the uniquenesses of the measured variables was not achieved with the data reported here.

Two Challenges for Future Research

Although tests of measurement invariance have become increasingly common, there are at least two major challenges facing this area of research.

The choice of a criterion for fit. The first challenge is the search for a criterion for judging the fit of invariance tests. Currently, the most frequently used standard is chi-square difference statistic. Here, $\alpha = .05$ provides a well-established convention for statistical significance. However, the likelihood ratio (chi-square difference test) is sensitive to nonnormality[9] and has substantial power in large samples to detect small discrepancies between groups that may be of no theoretical or practice consequence (errors of approximation; see Browne & Cudeck, 1993). In addition, the likelihood ratio test assumes that the less restricted model is properly specified. Consequently, fit indexes have been suggested as providing an improved criterion. However, fit indexes do not have known sampling distributions so that significance tests are not possible. There is also currently a lack of complete understanding of how goodness-of-fit indexes change as invariance constraints are imposed in multigroup analyses.

In early work on this issue, McGaw and Jöreskog (1971) and Tucker and Lewis (1973) compared factor solutions of often-analyzed data sets. McGaw and Jöreskog concluded that a difference between models in the Tucker–Lewis Index (ρ) of less than .022 indicated a negligible difference in fit, whereas Tucker and Lewis concluded this value should be .05. Cheung and Rensvold (2002), in an extensive simulation study of measurement invariance testing in first-order models,[10] exam-

[9]An alternative scaled chi-square difference test (Satorra & Bentler, 2001) can be used to minimize the effect of nonnormal data.

[10]We were unable to locate any study to date that has examined the performance of fit indexes in measurement invariance testing in second-order models.

ined the properties of 20 goodness-of-fit indexes in models containing unimportant small errors of approximation. They found that several commonly used indexes like χ^2/df and the change in the Tucker–Lewis Index showed poor performance in testing for measurement invariance. Cheung and Rensvold concluded that changes in CFI, Gamma hat, and McDonald's Noncentrality Index provided the best performance in terms of not being overly sensitive to small errors of approximation. However, as with all simulation work, the conclusions cannot be generalized beyond the set of conditions investigated in the study.

Perhaps of most practical importance, Cheung and Rensvold (2002) did not investigate the power of the fit indexes to reject the null hypothesis of invariance given that there is a failure of an invariance constraint that is of sufficient magnitude to be of material importance. For example, fit indexes were originally developed for covariance structure models in which there is no mean structure. Existing fit indexes appear to be differentially sensitive to differences in mean structure. A possible limitation of the CFI is that some work suggests that it is relatively insensitive to mean structure so that important differences in the intercepts of measured and latent variables may not be adequately detected using this fit index. Such problems may call for the development and evaluation of comparisons of new fit indexes that are more focused on the effects of specific constraints (e.g., threshold invariance) that are being imposed.

Clearly, additional work will be required to confirm or amend Cheung and Rensvold's (2002) suggestions. At present, their recommended guidelines might be treated cautiously as a possibly liberal test (i.e., running the risk of having insufficent power to detect invariance), whereas the likelihood ratio test might be treated as a too-conservative test (i.e., running the risk of being too likely to detect invariance where no appreciable invariance exists). In cases in which the two approaches disagree, analysts should note this in their reports and present their arguments for and against measurement invariance. As Bollen and Long (1983) noted, "The test statistics and fit indices are very beneficial, but they are no replacement for sound judgment and substantive expertise" (p. 8).

Strategies when measurement invariance fails. A second challenge is what strategies researchers should take when tests of measurement invariance fail. In early stages of measurement development, it may be possible to discard problematic items. These items can be identified in an initial study and discarded, with the remaining items ideally showing evidence of measurement invariance in a replication study. This strategy assumes that it is possible to identify which items are not invariant, often a difficult task, and that there are sufficient items remaining to properly specify each of the hypothesized factors of interest. An alternative approach that has been proposed is partial measurement invariance (Byrne, Shavelson, & Muthén, 1989). In a partial measurement invariance model, the in-

variant items are constrained to be equal, whereas the noninvariant items are allowed to differ across the groups. Once again, it is often difficult to identify the specific items that are not invariant and, more important, the consequences of partial invariance are largely unknown. Studies are only beginning to examine how partial measurement invariance affects group comparisons in terms of the relation between the factors, the factor means, and the prediction of external criteria (Millsap & Kwok, 2004). Finally, generalizations of work on measurement invariance within the item response theory framework may provide potential solutions (see, e.g., Embretson & Reise, 2000, chapter 10, for an introduction). For example, when measurement invariance fails at the intercept level, it may be possible to adjust for differences in the intercepts so that the means of latent variables can be properly compared. Once again, application of these promising procedures to measurement invariance in the CFA context is at a relatively early stage of development and further work is needed to evaluate their performance.

Conclusion

Despite remaining challenges, testing measurement invariance is important in answering many research questions. Establishing measurement invariance allows researchers to make appropriate comparisons between such important groups as males and females, different cultural groups, age groups, occupational groups, or even experimental and control groups. The work on testing measurement invariance in CFA models has made great progress since its initial development more than 30 years ago. More complete approaches involving both mean and covariance structure have been developed so that there is now a strong parallel to the other major approach to examining measurement invariance, item response theory. This article attempted to add to this development by providing an illustration for researchers of the procedures through which second-order factor models can be tested for measurement invariance.

ACKNOWLEDGMENTS

This project was supported by a National Institute of Nursing Research grant (NR04817) to Karen H. Sousa. We particularly thank Roger Millsap for his insights on testing measurement invariance and comments on earlier versions of this article. We also thank Oi-Man Kwok for his assistance and his comments on an earlier version of this article.

Fang Fang Chen is now at the Department of Psychology, University of Delaware.

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.

Alwin, D. F., & Jackson, D. J. (1981). Applications of simultaneous factor analysis to issues of factorial invariance. In D. Jackson & E. Borgotta (Eds.), *Factor analysis and measurement in sociological research: A multi-dimensional perspective* (pp. 68–119). Beverly Hills, CA: Sage.

Bentler, P. M. (1989). *EQS structural equations program manual.* Encino, CA: Multivariate Software.

Bentler, P. M. (1990). Comparative fit indices in structural equation models. *Psychological Bulletin, 107,* 238–246.

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

Berger, B. E., Ferrans, C., & Lashley, F. R. (2001). Measuring stigma in people with HIV: Psychometric assessment of the HIV Stigma Scale. *Research in Nursing and Health, 24,* 518–529.

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

Bollen, K. A., & Long, J. S. (Eds.). (1983). *Testing structural equation models.* Newbury Park, CA: Sage.

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

Byrne, B. M. (1995). Strategies in testing for an invariant second-order factor structure: A comparison of EQS and LISREL. *Structural Equation Modeling, 2,* 53–72.

Byrne, B. M., & Campbell, T. L. (1999). Cross-cultural comparisons and the presumption of equivalent measurement and theoretical structure: A look beneath the surface. *Journal of Cross-Cultural Psychology, 30,* 555–574.

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

Byrne, B. M., & Watkins, D. (2003). The issue of measurement invariance revisited. *Journal of Cross-Cultural Psychology, 34,* 155–175.

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

Cudeck, R. (1989). Analysis of correlation matrices using covariance structure models. *Psychological Bulletin, 105,* 317–327.

DeYoung, C. G., Peterson, J. B., & Higgins, D. M. (2002). Higher-order factors of the Big Five predict conformity: Are there neuroses of health? *Personality and Individual Differences, 33,* 533–552.

Eid, M., Lischetzke, T., Nussbeck, F., & Trierweiler, L. (2003). Separating trait effects from trait-specific method effects in multitrait–multimethod models: A multiple-indicator CT-C(M–1) Model. *Psychological Methods, 8,* 38–60.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Gotay, C. C., Blaine, D., Haynes, S., Holup, J., & Pagano, I. S. (2002). Assessment of quality of life in a multicultural cancer patient population. *Psychological Assessment, 14,* 439–448.

Gustafsson, J., & Balke, G. (1993). General and specific abilities as predictors of school achievement. *Multivariate Behavioral Research, 28,* 407–434.

Harlow, L. L., & Newcomb, M. D. (1990). Towards a general hierarchical model of meaning and satisfaction in life. *Multivariate Behavioral Research, 25,* 387–405.

Hills, P., & Argyle, M. (2002). The Oxford Happiness Questionnaire: A compact scale for the measurement of psychological well-being. *Personality and Individual Differences, 33,* 1071–1082.

Horn, J. L., McArdle, J. J., & Mason, R. (1983). When is invariance not invariant: A practical scientist's look at the ethereal concept of factor invariance. *Southern Psychologist, 4,* 179–188.

Hoyle, R. H. (1991). Evaluating measurement models in clinical research: Covariance structure analysis of latent variable models of self-conception. *Journal of Consulting and Clinical Psychology, 59,* 67–76.

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

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

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

Jöreskog, K. G., & Sörbom, D. (1999). *LISREL 8: User's reference guide* (2nd ed.). Chicago: Scientific Software International.

Judge, T. A., Erez, A., Bono, J. E., & Thoresen, C. J. (2002). Are measures of self-esteem, neuroticism, locus of control, and generalized self-efficacy indicators of a common core construct? *Journal of Personality and Social Psychology, 83,* 693–710.

Kim, S., Brody, G. H., & Murry, V. M. (2003). Factor structure of the early adolescent temperament questionnaire and measurement invariance across gender. *Journal of Early Adolescence, 23,* 268–294.

Leone, L., Perugini, M., Bagozzi, R. P., Pierro, A., & Mannetti, L. (2001). Construct validity and generalizability of the Carver-White Behavioural Inhibition System/Behavioural Activation System Scales. *European Journal of Personality, 15,* 373–390.

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

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

Marsh, H. W., Ellis, L. A., & Craven, R. G. (2002). How do preschool children feel about themselves? Unraveling measurement and multidimensional self-concept structure. *Developmental Psychology, 38,* 376–393.

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.

McArdle, J. J., & Nesselroade, J. R. (1994). Using multivariate data to structure developmental change. In S. H. Cohen & H. W. Reese (Eds.), *Life-span developmental psychology: Methodological contributions* (pp. 223–267). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

McGaw, B., & Jöreskog, K. G. (1971). Factorial invariance of ability measures in groups differing in intelligence and socioeconomic status. *British Journal of Mathematical and Statistical Psychology, 37,* 234–251.

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

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

Reise, S. P., Smith, L., & Furr, R. M. (2001). Invariance on the NEO PI–R neuroticism scale. *Multivariate Behavioral Research, 36,* 83–110.

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.

Rindskopf, D., & Rose, T. (1988). Some theory and applications of confirmatory second-order factor analysis. *Multivariate Behavioral Research, 23,* 51–67.

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test for moment structure analysis. *Psychometrika, 66,* 507–514.

Sörbom, D. (1978). An alternative to the methodology for analysis of covariance. *Psychometrika, 43,* 381–396.

Spearman, C. (1927). *The abilities of man.* London: Macmillan.

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

Stewart, A. L., Hays, R. D., & Ware, J. E. (1992). Health perceptions, energy/fatigue, and health distress measures. In A. L. Stewart & J. E. Ware (Eds.), *Measuring functioning and well-being: The medical outcome study approach* (pp. 143–172). Durham, NC: Duke University Press.

Stewart, A. L., & Ware, J. E. (1992). *Measuring functioning and well-being: The medical outcome study approach.* Durham, NC: Duke University Press.

Steyer, R., Ferring, D., & Schmitt, M. J. (1992). States and traits in psychological assessment. *European Journal of Psychological Assessment, 8,* 79–98.

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

Ullman, J. B. (2001). Structural equation modeling. In B. G. Tabachnick & L. S. Fidell (Eds.), *Using multivariate statistics* (pp. 653–771). Boston: Allyn & Bacon.

Van de Vijver, F. J. R., & Leung, K. (1997). Methods and data analysis for comparative research. In J. Berry, Y. Poortinga, & J. Pandey (Eds.), *Handbook of cross-cultural psychology* (Vol. 1, pp. 259–300). Boston: Allyn & Bacon.

Ware, J. E., Davies-Avery, A., & Brook, R. H. (1980). *Conceptualization and measurement of health for adults in the Health Insurance Study: Vol. 6. Analysis of relationships among health status measures.* Santa Monica, CA: RAND.

West, S. G., Finch, J. F., & Curran, P. J. (1995). Structural equation models with nonnormal variables: Problems and remedies. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 56–75). Thousand Oaks, CA: Sage.

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 281–324). Washington, DC: American Psychological Association.