Do Self-Report Instruments Allow Meaningful Comparisons Across Diverse Population Groups?

Testing Measurement Invariance Using the Confirmatory Factor Analysis Framework

Steven E. Gregorich, PhD

Abstract: Comparative public health research makes wide use of self-report instruments. For example, research identifying and explaining health disparities across demographic strata may seek to understand the health effects of patient attitudes or private behaviors. Such personal attributes are difficult or impossible to observe directly and are often best measured by self-reports. Defensible use of self-reports in quantitative comparative research requires not only that the measured constructs have the same meaning across groups, but also that group comparisons of sample estimates (eg, means and variances) reflect true group differences and are not contaminated by group-specific attributes that are unrelated to the construct of interest. Evidence for these desirable properties of measurement instruments can be established within the confirmatory factor analysis (CFA) framework; a nested hierarchy of hypotheses is tested that addresses the cross-group invariance of the instrument's psychometric properties. By name, these hypotheses include configural, metric (or pattern), strong (or scalar), and strict factorial invariance. The CFA model and each of these hypotheses are described in nontechnical language. A worked example and technical appendices are included.

Key Words: confirmatory factor analysis, measurement invariance, factorial invariance, metric invariance, scalar invariance, strict factorial invariance

(*Med Care* 2006;44: S78–S94)

1. INTRODUCTION

Fundamentally, comparative research seeks to identify similarities and differences as well as explain differences across known population groups. Comparative investigations are often quantitative, raising concerns about whether instrumentation provides a valid basis for making group comparisons.[1–4] When measurements are provided by self-report or other fallible methods, concerns about instrumentation are often exacerbated. This is especially true when measurements target attributes that are not directly observable such as attitudes and beliefs, intentions and motives, mood states, or behaviors that occur in private settings.

Often, such attributes are measured using multi-item self-report instruments. Each item is considered an imperfect measure of the attribute of interest but, as a whole, a set of similar items is hoped to provide valid indirect assessment of the targeted attribute. Responses to items that are believed to represent a single shared attribute are usually summed to form a composite measure, which under reasonable circumstances will be more reliable.[5] Because the attributes of interest are not directly observed, they are referred to as latent variables or, in the jargon of factor analysis, common factors.[6,7]

The confirmatory factor analysis (CFA) framework provides a means to test the construct validity of item sets, ie, whether item sets are indirect measures of hypothesized latent variables.[8] Furthermore, CFA can test whether evidence of construct validity is invariant across 2 or more population groups as well as whether group comparisons of sample estimates reflect true group differences and are not contaminated by group-specific attributes that are unrelated to the construct of interest.[1,2,9–15] Available tests form a nested hierarchy defining several forms of factorial invariance. This hierarchy of tests provides increasing evidence of measurement invariance.[2,12] The results of these tests help to determine which types of quantitative group comparisons are defensible.

The various factorial invariance hypotheses and the required CFA methodology to test them have been discussed in the literature.[1,2,9,11,15–19] Even so, factorial invariance hypotheses are tested relatively infrequently. When they are tested, investigators have predominantly focused on invariance of construct validity.[19] Applications testing whether comparisons of group means are defensible increasingly appear in the literature, but they are still rare and scattered across substantive domains.[19–28]

Why is factorial invariance testing relatively rare? One possibility is that many investigators lack the requisite technical skills, although several accessible articles explain the mechanics of factorial invariance testing.[1,15–19] Perhaps a

From the University of California, San Francisco.
Supported by grant no. P30 AG15272 to the University of California, San Francisco from the National Institutes on Aging.
Reprints: Steven E. Gregorich, University of California, San Francisco, 3333 California Street, Suite 335, San Francisco, CA 94143-0856. E-mail: gregorich@medicine.ucsf.edu.
Copyright © 2006 by Lippincott Williams & Wilkins
ISSN: 0025-7079/06/4400-0078

more fundamental possibility is a lack of awareness in the scientific community. Many investigators may not understand that lack of measurement invariance can take many forms, each with specific threats to substantive quantitative group comparisons, but again, these issues have been discussed in the articles cited previously. A related possibility is that many investigators do not have an intuitive understanding of *how* the various forms of factorial invariance logically defend specific quantitative group comparisons. In support of this possibility, most explanations of these logical underpinnings have been technical in nature. Meredith's landmark theoretical work was written in theorem/proof format.[2] Steenkamp and Baumgartner, although mathematically less rigorous than Meredith, referenced matrix-based regression equations when explaining the rationale underlying important factorial invariance concepts.[15] Other articles did not report tests of some important factorial invariance hypotheses and therefore did not discuss them.[16,18] Vandenberg and Lance provided a thorough review of the practice and mechanics of factorial invariance testing, but the authors did not explain how specific factorial invariance conditions support various substantive group comparisons.[19] This article addresses each of these possibilities but emphasizes nontechnical explanations intended to provide an intuitive understanding of the logic underlying the various factorial invariance hypotheses.

Section 2 includes a brief conceptual introduction to common factor models. In section 3, the primary forms of factorial invariance are described in nontechnical language. An example CFA analysis testing factorial invariance hypotheses as well as substantive group comparisons is presented in section 4. The discussion is in section 5. Technical details are included in 3 appendices.

2. CONCEPTUAL INTRODUCTION TO LATENT VARIABLES AND THE COMMON FACTOR MODEL

There are many definitions of latent variables reflecting different philosophical orientations.[29–31] In all definitions, latent variables are not directly observed and, in some cases, they are inherently unobservable (eg, mental constructs such as patient health beliefs). One type of latent variable is an unobserved causal mechanism (or common factor) that influences responses to a corresponding set of directly observed variables. Observed variables are also known as items or manifest variables.

Figure 1 graphically displays a generic 2-factor model of 4 items in 2 independent population groups (eg, blacks and whites). Common factor models represent a set of linear regression equations: each item is an outcome, the common factors are the explanatory variables, and there is one regression equation per item.[7,8,32–34] In Figure 1, the ovals represent common factors, rectangles represent items (ie, observed or manifest variables), and triangles represent means and intercepts ($\bar{x}$ and "1," respectively). Single-headed arrows represent values of regression parameters (ie, factor loadings; λ_{11}, λ_{12}, . . . λ_{42}), common factor means (κ_{11}, κ_{12}, κ_{21}, and κ_{22}), and intercepts (τ_{11}, τ_{12}, . . .τ_{42}). Double-headed arrows represent common factor variances (ϕ_{11}, ϕ_{12}, ϕ_{21}, and ϕ_{22}) and covariances ($\phi_{(1,2)1}$ and $\phi_{(1,2)2}$) as well as item residual variances (θ_{11}, θ_{12}, . . .θ_{42}). For each parameter, the first subscript indexes the common factor or item and the second subscript indexes group membership. Thus, λ_{32} is the factor loading for item 3 in group 2.

In this example, each item is associated with a single common factor. Such models are known as congeneric common factor models. The common factors have normal distributions by assumption and their univariate distributions are described by mean (eg, κ_{11}) and variance (eg, ϕ_{11}) parameters. Within each population group, the common factors can covary (eg, $\phi_{(1,2)1}$), but the common factors are uncorrelated across population groups because in this cross-sectional model, the groups are independent. As in ordinary bivariate linear regression, associated with each equation are a regression parameter (ie, factor loading), an intercept, and residuals with zero-mean and constant variation (eg, λ_{11}, τ_{11}, and θ_{11}, respectively, for item 1 in group 1). Conditional on the common factors, the items are uncorrelated. That is, the model specifies that within a population group, interitem correlations are fully explained by the common factors. Additional details are given in Appendix A.

What largely distinguishes the common factor model from a set of ordinary linear regression equations is that the explanatory variables—the common factors—are not directly observed. Because they are unobserved, common factors have no inherent scale of measurement; given the observed data, their means and variances are assigned by constraints placed on the model.[8,35] Often this includes constraining the value of one loading per common factor to equal unity and the corresponding item intercept to equal zero (eg, $\lambda_{11} = \lambda_{31} = \lambda_{12} = \lambda_{32} = 1$ and $\tau_{11} = \tau_{31} = \tau_{12} = \tau_{32} = 0$). Alternative model identifying constraints are possible. Importantly, however, although the scales of the common factors are arbitrary, under appropriate circumstances that are described subsequently, meaningful group differences in common factor means and variances can be estimated.

3. A HIERARCHY OF FACTORIAL INVARIANCE CONCEPTS

In this section, we discuss the primary forms of factorial invariance: dimensional, configural, metric (or pattern), strong factorial (or scalar), and strict factorial invariance.[2] These types of factorial invariance form a nested hierarchy primarily represented by increasing levels of cross-group equality constraints imposed on factor loading, item intercept, and residual variance parameters. Initially, each invariance concept is described in its "full" form. In a later section, the concept of partial invariance is addressed.

Cross-sectional factorial invariance concerns the invariance of corresponding parameters across independent population groups. Longitudinal factorial invariance concerns the invariance of corresponding parameters across time within a population group.[36] We only address cross-sectional factorial invariance.

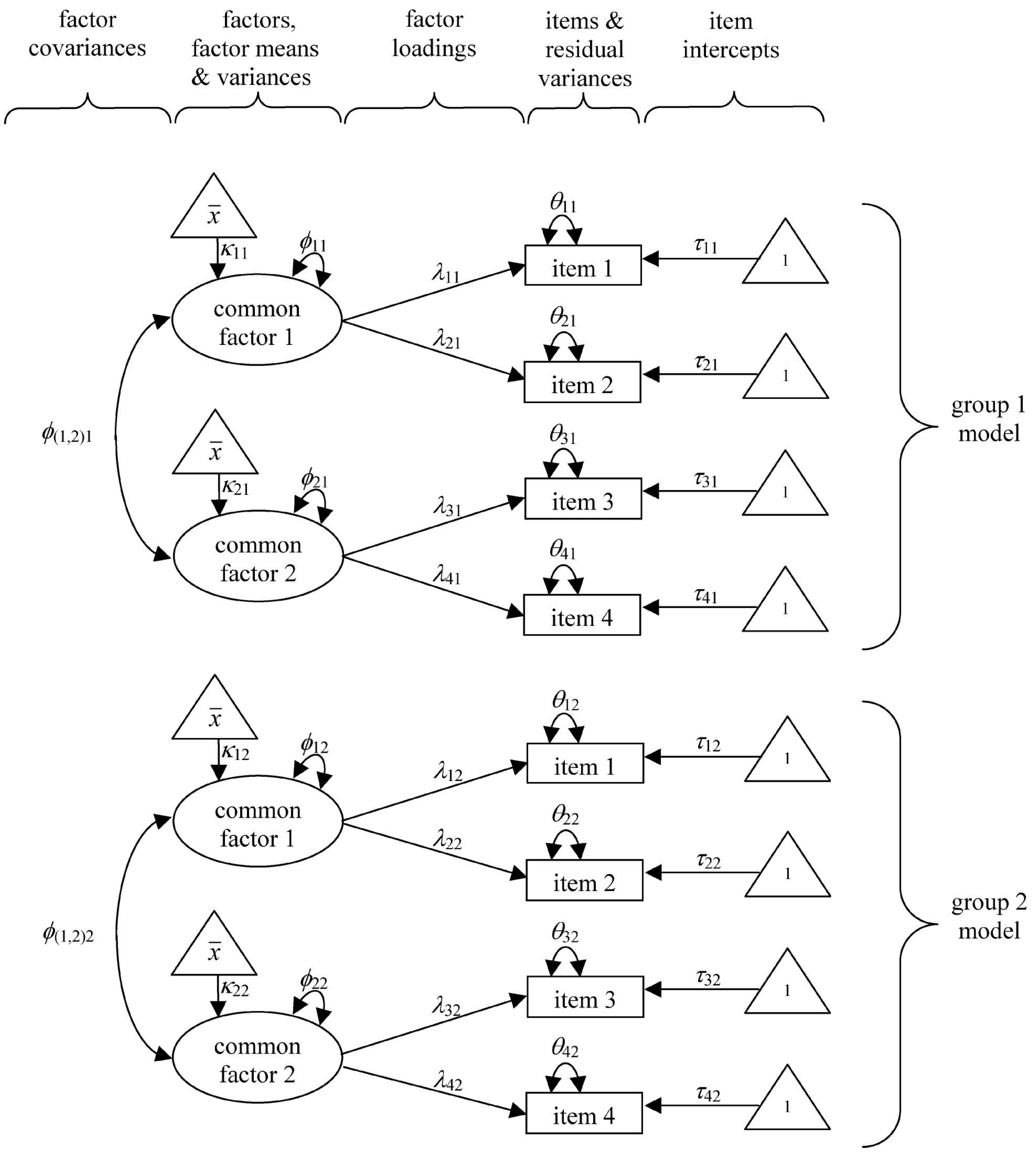

FIGURE 1. Generic 2-factor model in 2 groups.

Dimensional Invariance: Is the Same Number of Common Factors Present in Each Group?

Dimensional invariance simply requires that an instrument represents the same number of common factors across groups. Note that this form of invariance is only concerned with the number of factors in each group, not the specific configuration of items and factors. Clearly, instruments measuring the same number of factors across groups are desirable in quantitative comparative research. Logically, if an instrument represents differing numbers of factors across groups, then the meanings of one or more common factors must qualitatively differently across those groups.

The number-of-factors "problem" has a longstanding history in the psychometric literature. Within the exploratory factor analysis (EFA) framework, proposed solutions include those based on inspection of the eigenvalues of the item correlation matrix[37,38] and a χ^2 test of whether the number of extracted factors is sufficient (ie, whether the common factor model reproduces the item correlation matrix, within measurement error).[6,39–43] The same test can be performed within the CFA framework, but rarely is.[34,44] Regardless of the tools used for selecting the number of factors, the decision should not be based solely on empirical criteria. It is at least as important to consider theory, practical significance, parsimony, and past empirical results in the decision process.

The common factor model in Figure 1 is dimensionally invariant across groups. That is, the model for each group has 2 common factors. Dimensional invariance is desirable, but it does not provide evidence that quantitative group comparisons are defensible.

Configural Invariance: Are Common Factors Associated With the Same Items Across Groups?

Assuming dimensional invariance, configural invariance requires that each common factor is associated with identical item sets across groups. If an instrument measures the same common factors across groups, it logically follows that the item sets associated with each factor will be identical across those groups. Note that configural invariance does not require invariance of any parameter estimates across groups; it only requires that the item clusters are identical.

Traditionally, this form of invariance was more or less subjectively established by visually comparing rotated factor pattern matrices across group-specific EFA solutions; support for configural invariance was established when corresponding items had "substantial" loadings on the same common factor(s) across groups and "trivial" loadings on any other modeled factors. In CFA models, the item clusters are explicitly specified and a formal test of the configural invariance hypothesis is available. Therefore, CFA provides the superior framework for testing configural invariance. If a model with specified item clusters fits well in all groups, then configural invariance is supported. It is also worth noting that because no formal cross-group comparisons of parameter estimates are involved, any single-sample CFA replication of a previously supported factor model constitutes evidence of configural invariance.

The common factor model in Figure 1 is configurally invariant across groups. In each group, items 1 and 2 are associated with one common factor, whereas items 3 and 4 are associated with the other. Configural invariance is better than dimensional invariance, but it is not sufficient to defend quantitative group comparisons.

Metric Invariance: Do the Common Factors Have the Same Meanings Across Groups?

Assuming configural invariance, the metric invariance model requires corresponding factor loadings to be equal across groups. In Figure 1, this would require that $\lambda_{11} = \lambda_{12}$, and $\lambda_{21} = \lambda_{22}$ and $\lambda_{31} = \lambda_{32}$, and $\lambda_{41} = \lambda_{42}$. This level of invariance provides evidence that corresponding common factors have the same meaning across groups. This interpretation has intuitive appeal. After all, if common factors with identical meanings were operating in each group, we would expect identical relationships between the factors and the responses to a common set of relevant items. Conversely, it is difficult to imagine how metric invariance could hold if the common factors had different meanings across groups, especially if the factor loadings were substantial. Metric invariance is also known as pattern invariance stemming from the fact that the full set of factor loadings is known as the factor pattern matrix.

Within the EFA framework, ad hoc methods for testing metric invariance are based on correlating the estimated factor loadings across independent samples; the level of correlation indicates, more or less, the level of factor loading invariance across groups.[6] Multisample CFA methods are superior because they allow for simultaneous estimation of factor loadings across groups and provide a statistical test of the equal-loadings hypothesis. Metric invariance is tested by imposing equality constraints on corresponding factor loadings and fitting the factor model to sample data from each group simultaneously. If this model fits well, it suggests that the equality constraints are supported by the data. Metric invariance can also be tested by comparing the relative fit of 2 nested models: the fit of the configural and metric invariance models are compared and any significant worsening of fit suggests that the equal factor loadings hypothesis is not supported. Examples of both tests of metric invariance are provided in section 4.

If the metric invariance hypothesis is supported, then quantitative group comparisons of estimated factor variances and covariances are defensible. These comparisons are defensible because (1) the meanings of corresponding common factors are deemed invariant across groups and (2) the CFA model decomposes total item variation into estimated factor (ie, "true" score) and residual components. Therefore, group differences in common factor variation and covariation are not contaminated by possible group differences in residual variation. In contrast, metric invariance does not support group comparisons of observed (item or summed composite) variances and covariances, because observed measures confound common factor and residual variation. That is, assuming metric invariance, group differences in observed variation and covariation do not necessarily reflect group differences in common factor variation and covariation.

If metric invariance is not supported, then 2 interpretations are possible. The first is that one or more of the common factors, or at least a subset of the items, have different meanings across the population groups. A second possibility is that a subset of the factor loading estimates for one or more groups are biased resulting from so-called extreme response style (ERS), which can take 2 forms.[9,45] High ERS is a tendency to use extreme response options (eg, the "never" and "always" options of a 5-point ordered response set). This response style may occur in population groups whose members value decisiveness or certainty. Low ERS is the tendency to avoid extreme responses, favoring middling response options, and may be found in population groups that value humility and refraining from judgment.

ERS affects response variation; if ERS does not uniformly influence responses to all items within a population group, then item correlations and factor loading estimates will also be affected. Under some circumstances, the empirical distributions of item responses may strongly suggest ERS. More often, subjective judgment will guide decisions about whether a metric invariance hypothesis failed because of differential factor (or item) meanings or differential ERS across groups.

Strong Factorial Invariance: Are Comparisons of Group Means Meaningful?

When the metric invariance hypothesis is supported and the common factors are interpreted to have invariant meanings across population groups, another threat to measurement invariance should be considered: differential additive response bias, also known as differential acquiescence response styles.[9,45] Forces that are unrelated to the common factors such as cultural norms may systematically cause higher- or lower-valued item response in one population group compared with another. Unlike ERS, this response style is additive; it affects observed means but does not affect response variation. If these additive influences are not equivalent across groups, then they will contaminate estimates of group mean differences.

In the CFA model, item intercepts reflect these systematic, additive influences on responses to corresponding items that are constant in each group and are unrelated to the common factors. Assuming metric invariance, strong factorial invariance additionally requires that regressions of items onto their associated common factors yield a vector of intercept terms that is invariant across groups. In Figure 1, this would require that $\tau_{11} = \tau_{21}$, $\tau_{21} = \tau_{22}$, $\tau_{31} = \tau_{32}$, and $\tau_{41} = \tau_{42}$. Strong factorial invariance is also known as scalar invariance.

Evidence that corresponding factor loadings and item intercepts are invariant across groups suggests that (1) group differences in estimated factor means will be unbiased and (2) group differences in observed means will be directly related to group differences in factor means and will not be contaminated by differential additive response bias. To simplify matters further, it helps if we assume that all factor loadings are positive and the CFA model parameters have been rescaled so that the loadings associated with each common factor sum to unity within each group. The latter rescaling of model parameters is admissible because common factors are unobserved and have no natural scale of measurement. Given these assumptions, the expected group differences in composite means will equal the group differences in factor means; in this sense, when strong factorial invariance holds, we consider group differences in composite means to be unbiased estimates of group differences in corresponding factor means (Appendix A, equation A2). An essential point is that when corresponding item intercepts are invariant across groups they will cancel each other when group differences in observed means are estimated.

We present an example using a simple linear equation to clarify how strong factorial invariance can support defensible group mean comparisons. Suppose that mean observed patient weights are to be compared across 2 medical practices. The following equation relates mean observed patient weight to mean true body weight,

$$\text{mean observed weight} = \tau + \lambda \times \text{mean true weight}.$$

Here, τ is the intercept and λ is the parameter describing the relationship between true and observed body weight. By definition, the patient residual mean equals zero and drops out of the equation. Under optimal circumstances, the equation would equal,

$$\text{mean observed weight} = 0 + 1 \times \text{mean true weight},$$

where the intercept term equals zero and there is a one-to-one relationship between true and observed weight. In this case, observed weights equal true body weights. Real-life circumstances are rarely optimal. Suppose that in one practice, patients are weighed in their street clothes, but in the other practice, they wear examination gowns. Under these circumstances, the equations would differ across practices. Assuming all scales are accurate and that patient clothing and examination gowns average 1.2 and 0.2 kg, respectively, the resulting set of equations would be,

(practice 1)

$$\text{mean observed weight}_{p1} = 1.2 + 1 \times \text{mean true weight}_{p1}, \text{ and}$$

(practice 2)

$$\text{mean observed weight}_{p2} = 0.2 + 1 \times \text{mean true weight}_{p2}.$$

This example conceptually generalizes to the case in which metric invariance holds but strong factorial invariance does not. The mean true weights correspond to practice-specific common factor means, the parameter $\lambda = 1$ corresponds to an invariant factor loading, and the average weights of street clothes and examination gowns correspond to practice-specific intercept terms. (Note that strong factorial invariance can hold when $\lambda \neq 1$.) Although the scales used in each practice are equally valid instruments, procedural differences have introduced differential additive bias in weight measurements across practices. Consequently, differences in observed patient weights do not accurately reflect differences in true patient weights.

Consider possible implications of this differential additive bias. For mean true weights equaling 69 and 70 kg in practices 1 and 2, respectively, the mean observed weights would equal 70.2 in both practices. On the other hand, if the mean true weight equaled 70 kg in both practices, the mean observed weights would equal 71.2 and 70.2, respectively. Thus, differential additive bias can function to make observed mean differences smaller or larger than the true difference. This includes the possibility that no difference is observed when a real difference exists.

This example focused on differential additive bias introduced by procedural differences in taking weight measurements. Other sources of additive bias are possible. For example, norms within groups defined by race/ethnicity, culture, gender, age, period, or birth cohort may systematically act to raise or lower self-report item responses in ways that are unrelated to respondents' common factor scores. If such group-specific response tendencies exist, cross-sectional comparisons of observed means will be subject to differential additive response bias and are not readily interpretable—even if the measured common factors have the same meaning in each group.

Whereas tests of dimensional, configural, and metric invariance require fitting CFA models of covariance structures, testing strong factorial invariance requires modeling both mean and covariance structures. Mean and covariance structure models are less familiar to the general research community and this may partially explain why reported psychometric applications have tested strong factorial invariance relatively rarely. The strong factorial invariance model imposes equality constraints on corresponding factor loadings and item intercepts and fits the common factor model to sample data from each group simultaneously. Good fit suggests that the model constraints are consistent with the data. Strong factorial invariance can also be tested by comparing the fit of the metric and strong factorial invariance models. Any significant worsening of fit suggests that the equal item intercepts hypothesis is not supported.

Strict Factorial Invariance: Are Comparisons of Group Means and Observed Variances Defensible?

In typical applications, the common factor model decomposes the total variance of each item into 2 uncorrelated components: common factor and residual variation (Appendix A). If a goal is to compare observed variance estimates across population groups, the comparison should entirely reflect differences in common factor variation rather than being contaminated by differences in residual variation. Meaningful group comparisons of item or composite score variance estimates require an additional form of factorial invariance: residual invariance. Assuming metric (not strong factorial) invariance, residual invariance additionally requires that corresponding item residual variances are invariant across groups. When corresponding item residual variances are invariant, their effects will cancel in cross-group comparisons of observed variance estimates.

By itself, residual invariance does not support meaningful comparisons of group means. Because group mean comparisons are almost always of interest, residual invariance is of limited practical value. For defensible comparisons of both observed mean and variance estimates across population groups, evidence of strict factorial invariance should be obtained.[2] The strict factorial invariance model imposes cross-group equality constraints on corresponding factor loadings, item intercepts, and item residual variances. In other words, the strong factorial invariance model is further restricted so that corresponding item residual variances are invariant across groups. In Figure 1, this requires the following additional equality constraints: $\theta_{11} = \theta_{12}$, $\theta_{21} = \theta_{22}$, $\theta_{31} = \theta_{32}$, and $\theta_{41} = \theta_{42}$.

If the strict factorial invariance model holds, and the model parameters are scaled as described in the previous section, then expected group differences in composite means and variances will equal corresponding group differences in factor means and variances (Appendix A, equations A2 and A4). In summary, when corresponding item intercepts and residual variances are invariant across groups, they will cancel each other when group differences in observed means and variances are estimated.

TABLE 1. Summary of Factorial Invariance Testing

Invariance Hypothesis	Invariance Test	Defensible Substantive Quantitative Group Comparisons
Dimensional	Invariant no. of common factors	None
Configural	+ Invariant item/factor clusters	None
Metric (pattern)	+ Invariant factor loadings	Factor variances/ covariances
Strong factorial (scalar)	+ Invariant item intercepts	+ Factor and observed means
Strict factorial	+ Invariant residual variances	+ Observed variances/ covariances

Summary of Factorial Invariance

Table 1 summarizes the factorial invariance hypotheses as well as associated cross-group constraints and defensible substantive quantitative group comparisons. Group comparisons of observed or factor means, variances, and covariances are substantive in nature; these comparisons do not test measurement invariance hypotheses.[2,46,47] Instruments should be sensitive to group differences in construct means and variation. Examples of testing group differences in factor and observed means and variances are provided in section 4.

Partial Factorial Invariance

So far, we have discussed only the "full" version of each form of factorial invariance (eg, for metric invariance, *all* corresponding factor loadings are invariant across groups). Any one of the "full" forms of invariance described previously can be relaxed to obtain partial factorial invariance (eg, partial configural invariance, partial strong factorial invariance). As an example, in a one-factor model with 4 items, 3 of 4 factor loadings may be invariant, whereas the fourth differs across groups; this is known as partial metric invariance. Byrne et al first formally described partial invariance within the CFA framework.[16] Steenkamp and Baumgartner developed the concepts further.[15] According to those authors, a finding of partial invariance suggests that the substantive group comparisons associated with the corresponding "full" invariance hypotheses, as summarized in Table 1, are defensible. Essentially, only the subset of items meeting the metric, strong, or strict factorial invariance criteria are used to estimate associated group differences. For example, in a partial strong factorial invariance model, only those items meeting strong factorial invariance criteria actually contribute to estimates of group differences in factor means[15]; if observed means are to be compared, only those items meeting strong factorial invariance criteria are included in composite measures.

4. APPLICATION

Data

The example data are responses to the Center for Epidemiologic Studies Depression scale (CES-D)[48] collected as part of the National Health and Nutrition Examination Survey (NHANES) 1982–1984 Epidemiological Follow-up.[49,50] The CES-D is a self-report 20-item measure of depressive symptoms. Respondents indicate the degree to which they experienced each symptom during the prior week using 4 ordered response options ("0" rarely or none of the time, less than 1 day; "1" some or a little of the time, 1–2 days; "2" occasionally or a moderate amount of time, 3–4 days; "3" most or all of the time, 5–7 days). The CES-D is commonly believed to measure 4 factors. We only consider items from the Somatic and Retarded Activity factor with factor loadings greater than 0.40 in Radloff's original report[48] (items 1, 2, 7, 11, and 20; Appendix B) and the 248 black and 2004 white men over age 50 with complete data on these 5 items.

Modeling Approach

Multisample single-factor CFA models were fit to the data from black and white men using maximum likelihood (ML) estimation with LISREL 8.54[51](Fig. 2). To identify the CFA model, one factor loading per racial/ethnic group was fixed to unity and the corresponding intercept set to zero. Item responses were treated as continuously distributed. The empirical item distributions were nonnormal, which was expected to yield inflated ML χ^2 goodness-of-fit test statistics and underestimated parameter standard errors.[52–54] Therefore, the Satorra-Bentler (SB) scaled χ^2 and robust parameter standard errors were estimated, which help correct these biases.[55] The clustered sampling design of the NHANES was not modeled. However, the scaled χ^2 and robust standard errors mitigated the impact of intracluster response dependency on goodness-of-fit tests and standard error estimates.[56] Sampling weights were not available in the public use data.

Multisample CFA models fit to the data included the configural, metric, strong, partial strong, and partial strict factorial invariance models as well as the equal factor means and equal factor variances models. In addition to testing the overall fit of each model, we tested the relative fit of nested models; the difference in ML χ^2 statistics ($\Delta\chi^2$) and difference in model degrees of freedom were computed and compared with a central χ^2 distribution. In each case, the χ^2 difference tested whether the more constrained model (ie, that model imposing more cross-group equality constraints) resulted in a significant worsening of fit. To simplify presentation, corresponding comparisons of SB scaled χ^2 test statistics were not reported.[57] Additional fit indices included the root mean square error of approximation (RMSEA),[58] the expected cross-validation index (ECVI),[59] and the comparative fit index (CFI).[60] RMSEA values below 0.05 and CFI values above 0.95 are commonly considered rough indicators of approximate model fit.[59,61] In a set of competing models fit to the same data, relatively lower ECVI values suggest models that are more likely to replicate in independent samples of the same size.

When exploring partial factorial invariance models, 2 approaches to empirical model modification were considered. First, the traditional approach was used: consult so-called modification indices (or Lagrange multiplier tests) to determine which cross-group equality constraint, if any, most significantly contributed to lack of fit; free that constraint; reestimate the model; and reiterate this process as needed. Second, the method of Cheung and Rensvold was considered, which tests the cross-group invariance of parameters associated with each pair of items.[17] As an example, a one-factor model with 3 items has 3 within-group pairs of factor loadings: those for items 1 and 2, items 1 and 3, as well as items 2 and 3. In this example, 3 separate CFA models test the cross-group invariance of the factor loadings associated with each pair of items; in each model, the factor loadings for the remaining item would be freely estimated within each group. Afterward, test results are examined to determine whether one or more (possibly overlapping) item sets emerge with evidence of invariant factor loadings across groups. If such multiple item sets exist, the investigator may prefer one over the others; the choice of item set could be based on several criteria: the highest sum of squared factor loadings, the largest item set, the "best" face validity, and so on. Cheung and Rensvold[17] focused on partial metric invariance models, but their method naturally extends to investigation of partial strong and strict factorial invariance. The Cheung and Rensvold[17] approach is an improvement over the traditional method because it more easily allows identification of potential multiple item sets with invariant parameter estimates.

Tests of Factorial Invariance Hypotheses

Multisample CFA analyses tested the factorial invariance hypotheses. First, a one-factor model of mean and covariance structures was simultaneously fit to the data from black and white men. This model imposed no equality constraints on parameter estimates across samples. With a one-factor model, dimensional and configural invariance are identical. The model fit suggested reasonable evidence for dimensional/configural invariance, $SB\chi^2_{10} = 18.08$, $P = 0.04$ (Table 2). Furthermore, models fit separately for each racial/ethnic group produced a nonsignificant SB χ^2, suggesting the model fit well in both groups. The next model constrained corresponding factor loadings to be equal across groups and provided evidence of metric (or pattern) invariance, $SB\chi^2_{14} = 18.99$, nonsignificant (NS) and $\Delta\chi^2_4 = 0.45$, NS. The third model tested strong factorial (or scalar) invariance by additionally imposing equality constraints on corresponding item intercepts. This model was rejected, $SB\chi^2_{18} = 35.79$, $P < 0.01$ and $\Delta\chi^2_4 = 20.59$, $P < 0.001$. Modification indices suggested that the cross-group equality constraint on the intercept for item 7 ("effort") contributed most strongly to the lack of fit. The fourth model freely estimated this parameter in both groups and provided evidence of partial strong factorial invariance, $SB\chi^2_{17} = 23.64$, NS and $\Delta\chi^2_3 = 46.4$, NS. The fifth model tested partial strict factorial invariance by imposing additional cross-group equality constraints on all corresponding item residual variances except those for item 7. This model was also rejected, $SB\chi^2_{21} = 34.20$, $P < 0.05$ and

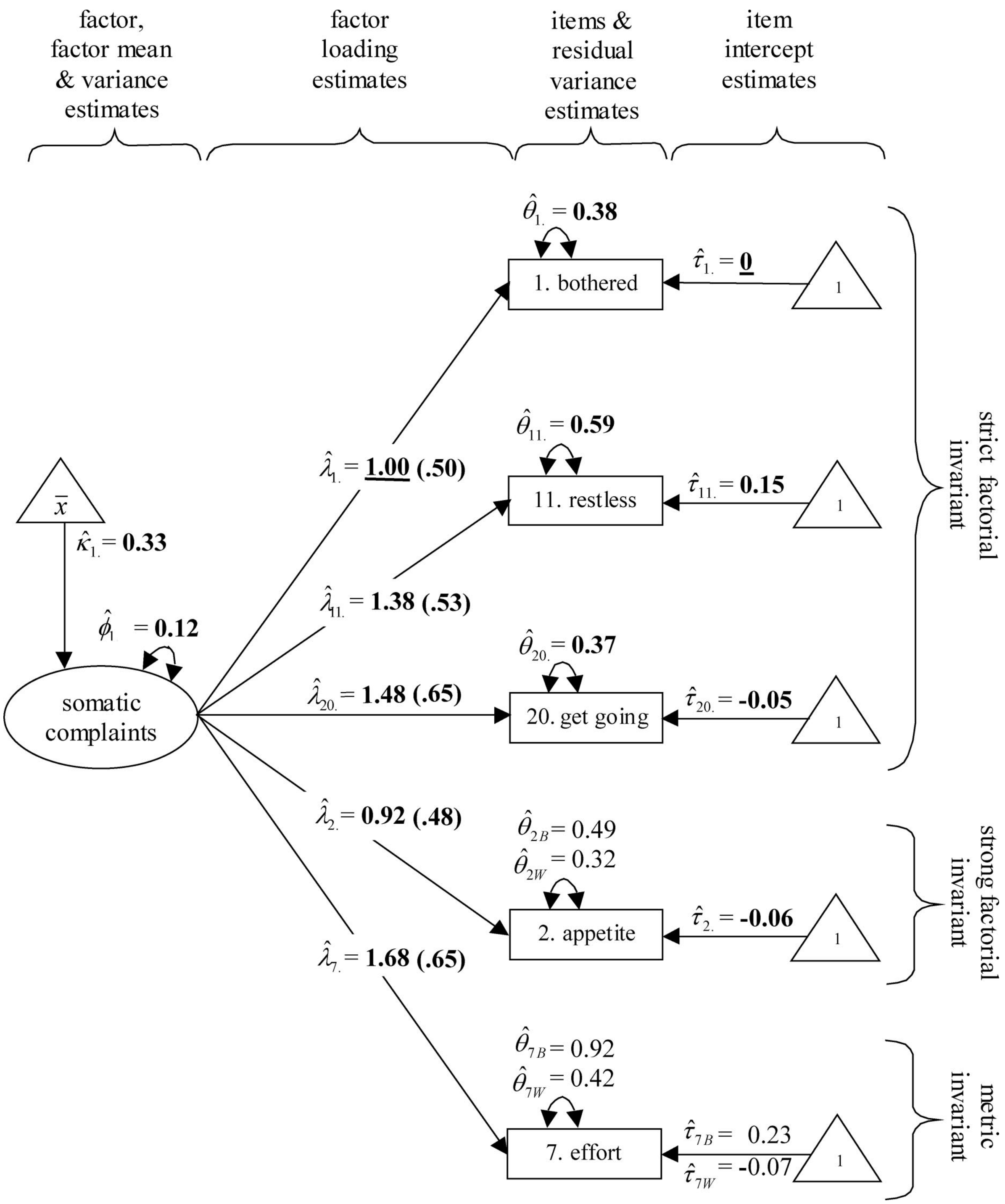
factor, factor mean & variance estimates
factor loading estimates
items & residual variance estimates
item intercept estimates
$\hat{\theta}_{1.} = \mathbf{0.38}$
1. bothered
$\hat{\tau}_{1.} = \underline{\mathbf{0}}$
1
$\hat{\theta}_{11.} = \mathbf{0.59}$
11. restless
$\hat{\tau}_{11.} = \mathbf{0.15}$
1
strict factorial invariant
$\bar{x}$
$\hat{\kappa}_{1.} = \mathbf{0.33}$
$\hat{\lambda}_{1.} = \underline{\mathbf{1.00}}$ **(.50)**
$\hat{\lambda}_{11.} = \mathbf{1.38}$ **(.53)**
$\hat{\phi}_{1} = \mathbf{0.12}$
$\hat{\theta}_{20.} = \mathbf{0.37}$
somatic complaints
$\hat{\lambda}_{20.} = \mathbf{1.48}$ **(.65)**
20. get going
$\hat{\tau}_{20.} = \mathbf{-0.05}$
1
$\hat{\lambda}_{2.} = \mathbf{0.92}$ **(.48)**
$\hat{\theta}_{2B} = 0.49$
$\hat{\theta}_{2W} = 0.32$
2. appetite
$\hat{\tau}_{2.} = \mathbf{-0.06}$
1
strong factorial invariant
$\hat{\lambda}_{7.} = \mathbf{1.68}$ **(.65)**
$\hat{\theta}_{7B} = 0.92$
$\hat{\theta}_{7W} = 0.42$
7. effort
$\hat{\tau}_{7B} = 0.23$
$\hat{\tau}_{7W} = \mathbf{-0.07}$
1
metric invariant

FIGURE 2. Partial strict factorial invariance model.

$\Delta\chi^2_4 = 18.62$, $P < 0.001$. The modification indices suggested freely estimating the residual variance for item 2 ("appetite"), resulting in a well-fitting partial strict factorial invariance model, $SB\chi^2_{20} = 21.68$, NS and $\Delta\chi^2_3 = 1.02$, NS. The method of Cheung and Rensvold[17] resulted in identical partial strong and partial strict factorial invariance models; LISREL code implementing the Cheung and Rensvold approach to assess partial strong factorial invariance is provided in Appendix C.

Comparisons of Factor Means and Variances

The partial strict invariance model was chosen as the basis of the models testing equality of factor means and variances. The seventh and eighth CFA models suggested evidence for equality

TABLE 2. Model Fit Summary

Model	SB χ^2	*df*	*P*	ML χ^2	Reference Model #	$\Delta\chi^2$	Δdf	ΔP	RMSEA	ECVI	CFI
1. Dimensional/configural	18.80	10	0.043	25.07					0.028	0.035	0.992
2. Metric	18.99	14	0.165	25.52	1	0.45	4	0.978	0.018	0.032	0.995
3. Strong	35.79	18	0.008	46.11	2	20.59	4	<0.001	0.030	0.035	0.983
4. Partial strong (τ_7 free)	23.64	17	0.130	30.16	2	4.64	3	0.200	0.019	0.031	0.994
5. Partial strict (τ_7 and θ_7 free)	34.20	21	0.034	48.78	4	18.62	4	<0.001	0.024	0.032	0.987
6. Partial strict (τ_7, θ_7, & θ_2 free)	21.68	20	0.358	31.18	4	1.02	3	0.796	0.009	0.027	0.998
7. Equal factor means ($\tau_{7,}$ θ_7, and θ_2 free)	22.30	21	0.382	31.68	6	0.50	1	0.480	0.007	0.027	1.00
8. Equal factor variances ($\tau_{7,}$ θ_7, & θ_2 free)	21.46	22	0.492	31.76	7	0.08	1	0.777	0	0.021	1.00

of factor means, $SB\chi^2_{21} = 22.30$, NS and $\Delta\chi^2_1 = 0.50$, NS, and factor variances, $SB\chi^2_{22} = 21.46$, NS and $\Delta\chi^2_1 = 0.08$, NS. Note that because its intercept was freely estimated in each group, item 7 did not contribute to the estimated group difference in factor means.[15] Finally, the RMSEA, ECVI, and CFI suggested that the equal factor means and equal factor variances models offered the best fit.

The final model is depicted in Figure 2. The parameter estimate subscripts identify item number and sample membership (ie, "B" or "W"). Parameter estimates constrained to be equal across samples have second subscripts equal to "·" and are in bold type. Standardized factor loadings are in parentheses. Model-identifying constraints are underlined.

Comparisons of Observed Means

The previously mentioned results suggested that all 5 CES-D items measured the same construct in both samples, ie, metric invariance was supported. However, metric invariance does not guarantee that group mean comparisons will be unbiased. To explore this further, we compared the observed means of all 5 items (Table 3). Black men had higher mean response levels on all items, except "restless." The largest difference was for "effort," but the previous analysis suggested this item was subject to differential additive response bias, invalidating group mean comparisons. We next computed composites by summing all 5 items as well as the 4 strong factorial invariant items (ie, removing "effort"). The 5-item composite suggested that black men reported significantly higher levels of somatic symptoms than did white men. However, the 4-item composite suggested no significant group difference and this result corresponded to the findings from the seventh CFA model, which tested equality of factor means across groups.

Comparisons of Observed Variances

Comparisons of item variances suggested no group differences among the 3 strict factorial invariant items. The remaining items "appetite" and "effort" had higher levels of response variation in the black sample, but the CFA models suggested group comparisons of observed response variation for those items would contaminate common and residual variation, essentially leading to meaningless comparisons. Variances of corresponding 5- and 3-item composites were also compared. In neither case did the group composite variances significantly differ. However, the group difference was larger for the 5-item composite; for the 3-item composite group levels in response variation were almost identical. This mirrored the result from the eighth CFA model, which tested the equality of factor variances across groups.

TABLE 3. Comparison of Observed Means and Variances

		Observed Means			Observed Variances		
Items	Equality Constraints	Black	White	Δ, *P*	Black	White	Ratio (*F'*), *P*
1. Bothered	Strict: λ,τ,θ	0.351	0.328	0.023, 0.64	0.496	0.498	1.00, 0.98
11. Restless	Strict: λ,τ,θ	0.589	0.612	−0.012, 0.70	0.874	0.823	1.06, 0.50
20. Get going	Strict: λ,τ,θ	0.456	0.437	0.019, 0.73	0.662	0.631	1.05, 0.59
2. Appetite	Strong: λ,τ	0.347	0.240	0.107, <0.02	0.576	0.427	1.35, <0.001
7. Effort	Metric: λ	0.807	0.486	0.321, <0.001	1.290	0.765	1.69, <0.001
Composites*							
5-item	At least metric	2.55	2.10	0.445, <0.01	7.99	7.18	1.11, 0.25
4-item	At least strong	1.74	1.62	0.124, 0.38	4.65	4.47	1.04, 0.68
3-item	Strict	1.40	1.38	0.017, 0.88	3.22	3.13	1.03, 0.75

*Five-item, sum of all 5 metric invariant items; 4-item, sum of the strong factorial invariant items (1, 2, 11, 20); 3-item, sum of the strict factorial invariant items (1, 11, and 20).

5. DISCUSSION

Quantitative comparative research places higher demands on instrumentation than single-group research. Single-group research can fare well with valid and reliable measures. Comparative research requires that instruments measure constructs with the same meaning across groups and allow defensible quantitative group comparisons. The example application demonstrated how measures can have the same meaning across population groups, but do not necessarily support defensible quantitative group comparisons. A group comparison of the 5-item composite measure suggested that black men reported significantly higher levels of somatic symptoms. However, the evidence suggested that this apparent racial disparity was spurious and was largely the result of differential additive response bias affecting the "effort" item.

From a quantitative comparative perspective, strict factorial invariance is universally desirable. Practical experience suggests that a finding of strong factorial (scalar) invariance is a more readily attainable goal that allows for defensible group comparisons of both observed and factor means as well as factor variances and covariances. Instruments that, at minimum, do not demonstrate partial strong factorial invariance may be counterproductive in comparative research. However, if focus is restricted to comparing the strength of corresponding regression parameters across groups, then metric invariance is sufficient if one of the following holds: (1) the regression model corrects for measurement errors in explanatory variables (eg, a structural equation model with latent variables) or (2) the reliabilities of corresponding explanatory variables are invariant across groups. (For a congeneric factor model, the combination of invariant factor loadings, residual variances, and factor variances is a sufficient, but not necessary, condition for equal item and composite reliability across groups. A less restrictive sufficient condition for equal item and composite reliability is metric invariance plus invariance of corresponding factor variance to residual variance ratios across groups.)

A common question concerns the limits of partial factorial invariance: Given a set of items hypothesized to represent a single common factor in 2 or more groups, how many items with invariant parameters are required to make valid quantitative group comparisons? There is little guidance on this matter. When faced with partial invariance, there are 3 broad options for how to proceed listed here by increasing conservativeness: (1) allow group comparisons on all items, regardless of any evidence for lack of measurement invariance; (2) restrict group comparisons to those items with appropriate invariant parameters; and (3) avoid group comparisons on any and all items identifying the factor. Millsap and Kwok[62] carefully considered how use of composite measures under option 1 affects selection. Focusing on the 2 more conservative options, there are algebraic and qualitative perspectives on this question, which are sometimes at odds. Option 2 represents an algebraic perspective in which items with varying loadings, intercepts, or residual variances are simply ignored; such items are treated as nuisances and are algebraically excluded from appropriate estimates of group differences. For example, a byproduct of the CFA partial strong factorial invariance model is that items with intercepts that vary across groups do not contribute to estimates of group differences in factor means.[15] If, on the other hand, observed means are compared, such items would be excluded from composite measures, effectively assigning them zero weight. From this perspective, at least 2 items with appropriate invariant parameters are required for defensible quantitative group comparisons. Technically, a single item with appropriate invariant parameters is sufficient for valid group comparisons. However, because of the required model-identifying constraints, tests of metric (and strong factorial) invariance require cross-group equality constraints on the factor loadings (and item intercepts) of at least 2 items. (Note that if the model-identifying constraints include fixing the factor loading and intercept for one item to equal unity and zero, respectively, in each group, then these represent cross-group equality constraints on the loadings and intercepts for that item.) Of course, larger numbers of items with invariant parameters will lead to more reliable estimates of group differences.

A qualitative perspective focuses on the meaning of common factors and, therefore, the invariance of factor loadings. From this perspective, any item with a factor loading that varies across population groups suggests that the factor has different meanings across those groups, and excluding such items from group comparisons ignores, but does not resolve, those qualitative differences. In its strict implementation, this perspective rejects quantitative group comparisons based on any and all items if at least one of them has varying factor loadings across groups (option 3).

The important contrast between the 2 perspectives is how they regard items with factor loadings that vary across groups. The algebraic perspective essentially treats such items as nuisances that can be ignored, whereas the qualitative perspective holds that such items signal qualitative group differences that render quantitative group comparisons as meaningless. Assuming at minimum that full metric invariance is supported, both the algebraic and qualitative perspectives allow items with varying intercepts or residual variances to be ignored in quantitative group comparisons, ie, under this circumstance, the 2 perspectives converge.

Some limitations of the example application are worth noting. We made reasonable methodological attempts to address the nonnormal distribution of the sample data and the clustered sampling design used to collect the data. Other methodological options can address these issues. For example, within the extended CFA framework, the item responses could have been modeled as ordinal rather than continuous.[63] A serious threat to the analyses reported in section 4 concerns the use of post hoc model modifications. Making such data-driven model modifications requires caution and invalidates probability values associated with subsequent χ^2 tests; the reported findings are provisional and require replication.[64,65] Nevertheless, the configural through the strong factorial invariance models were specified a priori.

The body of comparative research using self-report measures has been inadequately supported by appropriate psychometric evaluation. CFA applications testing factorial invariance hypotheses have typically focused on metric invariance, leaving open questions about the validity of quantitative group comparisons of means and observed variation. As an example, the psychometric properties of the CES-D have been studied relatively extensively. However, we know of no previously reported investigation testing any CES-D items for strong factorial invariance across black and white men. Other issues remain. Instruments that have been evaluated thoroughly in one or more population groups may eventually be administered in additional groups for which there is no supporting psychometric assessment. Furthermore, population groups are fluid. For example, results from NHANES data collected in 1984 do not necessarily generalize to black and white men sampled in 2006 or beyond. Period and cohort effects need to be considered. Statistical methodology has outpaced research practice in this arena.

ACKNOWLEDGMENTS

The author thanks Drs. Stuart Gansky, Roger Millsap, and Tor Neilands for comments on an earlier version of this manuscript.

REFERENCES

1. Little TD. Mean and covariance structures (MACS) analyses of cross-cultural data: practical and theoretical issues. *Multivariate Behav Res*. 1997;32:53–76.
2. Meredith W. Measurement invariance, factor analysis and factorial invariance. *Psychometrika*. 1993;58:525–543.
3. Millsap RE. Group differences in regression intercepts: implications for factorial invariance. *Multivariate Behav Res*. 1998;33:403–424.
4. Yoo B. Cross-group comparisons: a cautionary note. *Psychology and Marketing*. 2002;19:357–368.
5. Nunnally JC. *Psychometric Theory*. New York: McGraw Hill; 1978.
6. Gorsuch RL. *Factor Analysis*. Hillsdale, NJ: Lawrence Erlbaum Associates; 1983.
7. McDonald RP. *Test Theory: A Unified Treatment.* Hillsdale, NJ: Lawrence Erlbaum; 1999.
8. Bollen KA. *Structural Equations With Latent Variables*. New York: John Wiley and Sons; 1989.
9. Cheung GW, Rensvold RB. Assessing extreme and acquiescence response sets in cross-cultural research using structural equations modeling. *J Cross Cult Psychol*. 2000;31:187–212.
10. Jöreskog KG. Simultaneous factor analysis in several populations. *Psychometrika*. 1971;36:409–426.
11. Little TD. On the comparability of constructs in cross-cultural research: a critique of Cheung and Rensvold. *J Cross Cult Psychol*. 2000;31:213–219.
12. Meredith W, Teresi JA. An essay on measurement and factorial invariance. *Med Care*. 2006;44(Suppl 3):S69–S77.
13. Sörbom D. A general method for studying differences in factor mean and factor structures between groups. *Br J Math Stat Psychol*. 1974;27:229–239.
14. Sörbom D. Structural equation models with structured means. In: Jöreskog KG, Wold H, eds. *Systems Under Indirect Observation: Causality, Structure, and Prediction*. Amsterdam: North Holland; 1982:183–195.
15. Steenkamp J-BEM, Baumgartner H. Assessing measurement invariance in cross-national consumer research. *J Consum Res*. 1998;25:78–90.
16. Byrne BM, Shavelson RJ, Muthén B. Testing for the equivalence of factor covariance and mean structures: the issue of partial measurement invariance. *Psychol Bull*. 1989;105:456–466.
17. Cheung GW, Rensvold RB. Testing factorial invariance across groups: a reconceptualization and proposed new method. *J Manage*. 1999;25:1–27.
18. Horn JL, McArdle JJ. A practical and theoretical guide to measurement invariance in aging research. *Exp Aging Res*. 1992;18:117–144.
19. Vandenberg RJ, Lance CE. A review and synthesis of the measurement invariance literature: suggestions, practices, and recommendations for organizational research. *Organ Res Methods*. 2000;2:4–69.
20. Dolan CV. Investigating Spearman's hypothesis by means of multi-group confirmatory factor analysis. *Multivariate Behav Res*. 2000;35:21–50.
21. Hofer SM, Horn JL, Eber HW. A robust five-factor structure of the 16PF: strong evidence from independent rotation and confirmatory factorial invariance procedures. *Pers Individ Dif*. 1997;23:247–269.
22. Hong S, Malik ML, Lee MK. Testing configural, metric, scalar, and latent mean invariance across genders in sociotropy and autonomy using a non-Western sample. *Educ Psychol Meas*. 2003;63:636–654.
23. Kossowska M, Van Heil A, Chun WYK, et al. The need for cognitive closure scale: structure, cross-cultural invariance, and comparison of mean ratings between European-American and East Asian samples. *Psychol Belg*. 2002;42:267–286.
24. Marin BV, Tschann JM, Gómez CA, et al. Self-efficacy to use condoms in unmarried Latino adults. *Am J Community Psychol*. 1998;26:53–71.
25. Munet-Vilaró F, Gregorich SE, Folkman S. Factor structure of the Spanish version of the ways of coping questionnaire. *J Appl Soc Psychol*. 2002;32:1938–1954.
26. Pomplun M, Omar MH. The factorial invariance of a test of reading comprehension across groups of limited English proficiency students. *Appl Meas Educ*. 2001;14:261–283.
27. Shevlin M, Brunsden V, Miles JNV. Satisfaction with life scale: analysis of factorial invariance, mean structures, and reliability. *Pers Individ Dif*. 1998;25:911–916.
28. Whiteside-Mansell L, Corwyn RF. Mean and covariance structures analyses: an examination of the Rosenberg self-esteem scale among adolescents and adults. *Educ Psychol Meas*. 2003;63:163–173.
29. Bollen KA. Latent variables in psychology and the social sciences. *Annu Rev Psychol*. 2002;53:605–634.
30. Sobel ME. Causal inference in latent variable models. In: von Eye A, Clogg CC, eds. *Latent Variables Analysis: Applications for Developmental Research*. Thousand Oaks, CA: Sage; 1994:3–35.
31. Borsboom D, Mellenbergh GJ, Van Heerden J. The theoretical status of latent variables. *Psychol Rev*. 2003;110:203–219.
32. Bentler PM, Weeks DG. Interrelations among models for the analysis of moment structures. *Multivariate Behav Res*. 1979;14:169–185.
33. Bentler PM, Weeks DG. Linear structural equations with latent variables. *Psychometrika*. 1980;45:289–308.
34. Jöreskog KG. A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*. 1969;34:183–202.
35. MacCallum RC. Model specification: procedures, strategies, and related issues. In: Hoyle RH, ed. *Structural Equation Modeling: Concepts, Issues, and Applications*. Thousand Oaks, CA: Sage; 1995.
36. Meredith W, Horn J. The role of factorial invariance in modeling growth and change. In: Collins LM, Sayer AG, eds. *New Methods for the Analysis of Change*. Washington, DC: American Psychological Association; 2001.
37. Cattell RB. The screen test for the number of factors. *Multivariate Behav Res*. 1966;1:245–276.
38. Guttman L. Some necessary conditions for common factor analysis. *Psychometrika*. 1954;19:149–161.
39. Jöreskog KG. On the statistical treatment of residuals in factor analysis. *Psychometrika*. 1962;27:335–354.
40. Jöreskog KG. Some contributions to maximum likelihood factor analysis. *Psychometrika*. 1967;32:443–482.
41. Lawley DN. The estimation of factor loadings by the method of maximum likelihood. *Proceedings of the Royal Society of Edinburgh.* 1940:64–82.
42. Lawley DN, Maxwell AE. *Factor Analysis as a Statistical Method.* London: Butterworth; 1963.
43. Mulaik SA. *The Foundations of Factor Analysis*. New York: McGraw Hill; 1972.
44. Jöreskog KG. A general approach to confirmatory maximum likelihood factor analysis with addendum. In: Jöreskog K, Sörbom D, Magidson J, eds. *Advances in Factor Analysis and Structural Equation Models*. Cambridge, MA: Abt Books; 1979:21–43.
45. Baumgartner H, Steenkamp J-BEM. Response styles in marketing research: a cross-national investigation. *J Mark Res*. 2001;38:143–156.

46. Meredith W. Notes on factorial invariance. *Psychometrika*. 1964;29: 177–185.
47. Meredith W. Two wrongs may not make a right. *Multivariate Behav Res*. 1995;30:89–94.
48. Radloff LS. The CES-D scale: a self-report depression scale for research in the general populations. *Appl Psych Meas*. 1977;1:385–401.
49. Cornoni-Huntley J, Barbano HE, Brody JA, et al. National health and nutrition examination, I: epidemiologic followup survey. *Public Health Rep*. 1983;98:245–251.
50. Madans JH, Kleinman JC, Cox CS, et al. 10 years after NHANES I: report of initial followup, 1982–84. *Public Health Rep*. 1986;101:465–473.
51. Jöreskog KG, Sörbom D. Chicago: Scientific Software International; 2003.
52. Muthén BO, Kaplan D. A comparison of some methodologies for the factor analysis of non-normal Likert variables. *Br J Math Stat Psychol*. 1985;38:171–189.
53. Muthén BO, Kaplan D. A comparison of some methodologies for the factor analysis of non-normal Likert variables: a note on the size of the model. *Br J Math Stat Psychol*. 1992;45:19–30.
54. Curran PJ, West SG, Finch JF. The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychol Methods*. 1996;1:16–29.
55. Satorra A, Bentler PM. Corrections to test statistics and standard errors in covariance structure analysis. In: von Eye A, Clogg CC, eds. *Latent Variables Analysis: Applications for Developmental Research*. Thousand Oaks, CA: Sage Publications; 1994:399–419.
56. Muthén BO, Satorra A. Complex sample data in structural equation modeling. *Sociol Methodol*. 1995;25:267–316.
57. Satorra A, Bentler PM. A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*. 2001;66:507–514.
58. Steiger JH, Lind JM. *Statistically Based Tests for the Number of Common Factors*. Paper presented at the annual meeting of the Psychometric Society; Iowa City, IA; 1980.
59. Browne MW, Cudeck R. Alternative ways of assessing model fit. In: Bollen KA, Long JS, eds. *Testing Structural Equation Models*. Newbury Park, CA: Sage; 1993:136–162.
60. Bentler PM. Comparative fit indices in structural models. *Psychol Bull*. 1990;107:238–246.
61. Hu LT, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *SEM*. 1999;6:1–55.
62. Millsap RE, Kwok OM. Evaluating the impact of partial factorial invariance on selection in two populations. *Psychol Methods*. 2004;9: 93–115.
63. Muthén LK, Muthén B. *Mplus User's Guide*. Los Angeles: Muthén & Muthén; 2001.
64. Cliff N. Some cautions concerning the application of causal modeling methods. *Multivariate Behav Res*. 1983;18:81–105.
65. MacCallum R. Specification searches in covariance structure modeling. *Psychol Bull*. 1986;100:107–120.
66. Allen MJ, Yen WM. *Introduction to Measurement Theory*. Monterey, CA: Brooks & Cole; 1979.

APPENDIX A: STRONG AND STRICT FACTORIAL INVARIANCE

The Common Factor Model

The general common factor model in one group can be expressed as,

$$x_{ij} = \sum_{k=1}^{q} \lambda_{ik}\xi_{jk} + s_{ij} + \varepsilon_{ij} + \alpha_i;\ i = 1 \text{ to } p \text{ items};$$

$$j = 1 \text{ to } N \text{ respondents};\ k = 1 \text{ to } q \text{ factors};\ N > p > q,$$

where x_{ij} are the scores on p items or observed variables for each of N respondents, λ_{ik} are the regression coefficients or factor loadings of the items on each of the q common factors, ξ_{jk} are the common factor scores for each of the respondents, s_{ij} are the scores on the specific factors for each of the respondents, ε_{ij} are random errors of measurement, and α_1 are intercepts from the regression of each item onto the common factors. The model assumptions include (1) the ξ_k, s_i, ε_i are normally distributed, each with constant variance; (2) the ε_i are errors of measurement as defined by classical true score theory[66]; (3) the correlations among the ξ_k and s_i are zero; (4) the correlations among the s_i are zero. Meredith and Teresi describe further details.[12]

Throughout the remainder of this appendix, we consider a correctly specified single common factor model with positive factor loading values for all items (positive manifold). Because common factors are unobserved, they have no natural scale of measurement. Therefore, it is admissible to further assume that the parameters of the factor model have been rescaled so that the sum of the factor loadings within each group equals unity.

Mean Structure

We initially take an item-wise perspective. The means of the items, μ, can be expressed as $\mu_i = \lambda_i\kappa + (v_i + \alpha_i)$, where κ is the common factor mean and the v_i are the specific factor means. In most single-sample applications, only one mean parameter on the righthand side of the equation will be identified. Multisample applications allow estimation of additional mean parameters.[13,14] The parameters v_i and α_i are placed in parentheses because in typical multisample designs they cannot be estimated separately. By convention, we substitute τ_i for the quantity $(v_i + \alpha_i)$, resulting in $\mu_i = \lambda_i\kappa + \tau_i$.

A model with a single common factor decomposes the mean of each item into the product of a common factor mean and factor loading plus an intercept term. The population item mean vector is expressed as $\mathbf{M} = \boldsymbol{\Lambda}\kappa + \mathbf{T}$, where $\boldsymbol{\Lambda}$ is the vector of common factor loadings and $\mathbf{T}$ is the vector of item intercepts.

Covariance Structure

The variance of each item, σ_i^2, can be expressed as $\sigma_i^2 = \lambda_i^2\phi + (\upsilon_i + \omega_i)$, where ϕ is the common factor variance, υ_i are specific factor variances, and the ω_i are random error variances. The parameters υ_i and ω_i are placed within parentheses because in typical multisample designs, they cannot be estimated separately. For convenience, we substitute θ_i for the quantity $(\upsilon_i + \omega_i)$, yielding $\sigma_i^2 = \lambda_i^2\phi + \theta_i$. In this model, the variance of an observed variable is decomposed into 2 components: the product of the common factor variance and a squared factor loading plus residual variation. Consequently, for a one-factor model, the covariance structure of all items is expressed as $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\phi\boldsymbol{\Lambda}' + \boldsymbol{\Theta}$, or alternatively $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}'\phi + \boldsymbol{\Theta}$, where $\boldsymbol{\Theta}$ is a diagonal matrix of item residual variances.

Mean and Covariance Structures in Multisample Applications

In multisample applications, models are fit simultaneously to data from 2 or more samples. Questions naturally arise regarding the invariance of corresponding parameters across groups and the associated implications for comparative research. Here we consider a 2-group design and denote group membership by a subscript, *g*.

Comparing Observed Means Across Groups

The multigroup mean structure for a single-factor model can be written as, $\mathbf{M}_g = \mathbf{\Lambda}_g \kappa_g + \mathbf{T}_g$. When, at minimum, $\mathbf{\Lambda} = \mathbf{\Lambda}_1 = \mathbf{\Lambda}_2$ and $\mathbf{T} = \mathbf{T}_1 = \mathbf{T}_2$ (ie, strong factorial invariance), differences in item means will be proportional to the difference in common factor means,

$$\begin{aligned} \mathbf{M}_1 - \mathbf{M}_2 &= \mathbf{\Lambda}\kappa_1 + \mathbf{T} \ - (\mathbf{\Lambda}\kappa_2 + \mathbf{T}) \\ &= \mathbf{\Lambda}(\kappa_1 - \kappa_2). \end{aligned} \tag{A1}$$

Interest generally lies in computing composite scores for each respondent within in each group, c_{jg}, by summing observed scores on items associated with the common factor and assessing the composite mean group difference. Given a single common factor model, it follows from equation 1 that

$$\begin{aligned} E(\bar{c}_1 - \bar{c}_2) &= sum(\mathbf{M}_1 - \mathbf{M}_2) \\ &= sum(\mathbf{\Lambda})(\kappa_1 - \kappa_2) \\ &= \kappa_1 - \kappa_2, \end{aligned} \tag{A2}$$

where the $\bar{c}_g$ represent group composite means, *E*(.) is the expectation operator, *sum*(.) sums the elements of a matrix or vector, and $sum(\mathbf{\Lambda}) = 1$, by stated assumption, and therefore drops out of the equation. In this case, the expected difference in group composite means equals the difference in common factor means.

Comparing Observed Variation Across Groups

Assuming a single common factor model, the item covariance structures in multiple groups can be expressed as, $\Sigma_g = \mathbf{\Lambda}_g \, \mathbf{\Lambda}_g{}' \, \phi_g + \boldsymbol{\theta}_g$. When both $\mathbf{\Lambda} = \mathbf{\Lambda}_1 = \mathbf{\Lambda}_2$ and $\boldsymbol{\theta} = \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$ (ie, either residual or strict factorial invariance holds), it follows that group differences in item variation and covariation are proportional to group differences in common factor variation. In this case, a comparison of item covariance matrices across 2 groups can be expressed as

$$\begin{aligned} \mathbf{\Sigma}_1 - \mathbf{\Sigma}_2 &= \mathbf{\Lambda}\mathbf{\Lambda}' \, \phi_1 + \ \boldsymbol{\theta} - (\mathbf{\Lambda}\mathbf{\Lambda}' \, \phi_2 + \boldsymbol{\theta}) \\ &= \ \mathbf{\Lambda}\mathbf{\Lambda}' \, (\phi_1 - \phi_2). \end{aligned} \tag{A3}$$

Again, interest generally lies in computing composites by summing item scores and assessing group differences. If, at minimum, residual invariance holds, then the expected group difference in composite variation is expressed as,

$$\begin{aligned} E(\hat{\sigma}_1^2 - \hat{\sigma}_2^2) &= sum(\mathbf{\Sigma}_1 \ - \mathbf{\Sigma}_2) \\ &= sum(\mathbf{\Lambda}\mathbf{\Lambda}')(\phi_1 - \phi_2) \\ &= sum(\mathbf{\Lambda})^2(\phi_1 - \phi_2) \\ &= \phi_1 \ - \ \phi_2, \end{aligned} \tag{A4}$$

where $\hat{\sigma}_g^2$ is a group-specific composite variance estimate and $sum(\mathbf{\Lambda})^2 = 1$, by stated assumption. Thus, under residual or strict factorial invariance and the stated assumptions, the expected value of observed group differences in composite variation equals the group difference in common factor variation.

APPENDIX B: LISREL CODE FOR TESTING FACTORIAL INVARIANCE HYPOTHESES

```
Model for Black Men -----------------------------------------------------
!!-----------------------------------------------------------------------
!! The following LISREL code will fit the models described in the article.
!! The weight matrices required for estimation of the Satorra-Bentler
!! scaled chi-square and robust standard errors are not provided here.
!! Subsequently, only the ML chi-squares and model degrees of freedom
!! will match those reported in Table 1.
!!
!! Below, there are two sets of stacked LISREL code. The first set
!! (DA line through OU line) describes the model for Black men.
!! The second set describes the model for White men.  The invariance
!! hypotheses are tested by imposing equality constraints on parameter
!! estimates across the models for Black and White men.  The equality
!! constraints are specified in the second set of LISREL code.  On the
!! 'MO' line, the code '=PS' specifies that the corresponding parameters
!! are freely estimated in each group; the code '=IN' constrains the
!! corresponding parameters to be invariant across groups.
!!
!! All text following an exclamation mark ('!') is considered a comment
!! by LISREL.  Lines beginning with double exclamation marks ('!!')
!! provide documentation and should not be altered.  Lines beginning with
!! a single exclamation mark represent code that is selectively specified.
!! Users can comment and un-comment the appropriate lines of code below
!! to estimate each model.
!!
!! LISREL commands and keywords appear in uppercase. LISREL only reads
!! the first two characters of each command and keyword. The additional
!! characters following the first two capitalized letters are provided
!! for additional clarification and are not required.
!!
!! Items from the CES-D Somatic and Retarded Activity factor[48]
!! cesd01. I was bothered by things that usually don't bother me.
!! cesd02. I did not feel like eating; my appetite was poor.
!! cesd07. I felt that everything I did was an effort
!! cesd11. My sleep was restless
!! cesd20. I could not get "going."
!!
!! Data are from the NHANES 1982-1984 Epidemiological Follow-Up[49, 50]
!!-----------------------------------------------------------------------

DA_DATA  NI_#ITEMS=5  NO_#OBSERVATIONS=248  MA_MATRIX_TYPE=CM  NG_#GROUPS=2

LA_LABELS
cesd01 cesd02 cesd07 cesd11 cesd20

CM_COVARIANCE_MATRIX
0.49587
0.16935    0.57562
0.19368    0.17265    1.29030
0.14082    0.13106    0.31683    0.87469
0.18769    0.12476    0.34361    0.26508    0.66199

ME_MEANS
0.35081    0.34677    0.80645    0.58871    0.45565
```

```
MO_MODEL                               C !! 'C' continues to the next line
  NX_#ITEMS                  = 5       C !! 5 items in the model
  NK_#FACTORS                = 1       C !! 1 factor
  LX_FACTOR_LOADING_MATRIX   = FU,FR   C !! freely estimated
  TX_ITEM_INTERCEPTS         = FU,FR   C !! freely estimated
  TD_ITEM_RESIDUAL_VARIANCES = DI,FR   C !! freely estimated
  PH_FACTOR_VARIANCE         = SY,FR   C !! freely estimated
  KA_FACTOR_MEAN             = FU,FR     !! freely estimated

ST_START_VALUES   1.0 LX 1 1             !! needed to identify the model
FI_FIX_PARAMETERS     LX 1 1             !! needed to identify the model
FI_FIX_PARAMETERS     TX 1               !! needed to identify the model

OU_OUTPUT  ME_ESTIMATION_METHOD=ML


Model for White Men ---------------------------------------------------
DA NI=5 NO=2004 MA=CM

LA
cesd01 cesd02 cesd07 cesd11 cesd20

CM
0.49822
0.12634    0.42714
0.20046    0.17796    0.76513
0.18179    0.17049    0.27322    0.82264
0.16644    0.16163    0.31836    0.24146    0.63059

ME
0.32834    0.24002    0.48553    0.61228    0.43713

!! Model 1. Configural invariance model
MO  NX=5  NK=1  LX=PS  TX=PS  TD=PS  KA=PS  PH=PS

!! Model 2. Metric invariance model
!MO  NX=5  NK=1  LX=IN  TX=PS  TD=PS  KA=PS  PH=PS

!! Model 3. Strong factorial invariance model
!MO  NX=5  NK=1  LX=IN  TX=IN  TD=PS  KA=PS  PH=PS

!! Model 4. Partial strong factorial invariance model
!! free item intercept for the 3rd item: cesd07 'effort'
!MO  NX=5  NK=1  LX=IN  TX=IN  TD=PS  KA=PS  PH=PS
!FR_FREE_PARAMETER TX 3

!! Model 5. Partial strict factorial invariance model
!! free item intercept and residual variance for cesd07 'effort'
!MO  NX=5  NK=1  LX=IN  TX=IN  TD=IN  KA=PS  PH=PS
!FR TX 3
!FR TD 3 3

!! Model 6. Partial strict factorial invariance model
!! additionally free item residual variance of 2nd item: cesd02 'appetite'
!MO  NX=5  NK=1  LX=IN  TX=IN  TD=IN  KA=PS  PH=PS
!FR TX 3

!FR TD 3 3
!FR TD 2 2

!! Model 7. Equal factor means model
!MO  NX=5  NK=1  LX=IN  TX=IN  TD=IN  KA=IN  PH=PS
!FR TX 3
!FR TD 3 3
!FR TD 2 2

!! Model 8. Equal factor means and variances model
!MO  NX=5  NK=1  LX=IN  TX=IN  TD=IN  KA=IN  PH=IN
!FR TX 3
!FR TD 3 3
!FR TD 2 2

OU ME=ML
```

APPENDIX C: LISREL CODE IMPLEMENTING THE CHEUNG AND RENSVOLD[17] METHOD TO ASSESS PARTIAL STRONG FACTORIAL INVARIANCE

```
Model for Black men -----------------------------------------------------
!!-----------------------------------------------------------------------
!! The following LISREL code will implement the method of Cheung & Rensvold[17]
!! for testing partial strong factorial invariance.  This code can be
!! modified to test for partial metric or partial strict factorial
!! invariance.
!!
!! To simplify implementation of the Cheung & Rensvold approach, the model-
!! identifying constraints are different than those used in Appendix B. Here,
!! they include (1) the factor mean and variance are fixed to 0 and 1,
!! respectively, in the model for Black men, but freely estimated in the
!! model for White men and (2) cross-group equality constraints for at
!! least one factor loading and one item intercept estimate.
!!
!! The weight matrices required for estimation of the Satorra-Bentler
!! scaled chi-square and robust standard errors are not provided.
!!
!! When testing partial strong factorial invariance, the Cheung & Rensvold
!! approach tests the cross-group invariance of intercepts associated with
!! each possible pair of items (assuming, in this case, full metric
!! invariance). With five items there are 10 possible item pairs, resulting
!! in 10 tests. Because this procedure conducts multiple tests, a Bonferroni
!! or similar correction is recommended.  E.g., a Type-I error rate of 5%
!! suggests a corrected criterion p-value equal to .005 (i.e., .05/10).
!!
!! The following table summarizes ML chi-square p-values from the 10 tests,
!! obtained by executing the following LISREL code. For example, imposing
!! cross-group equality constraints on the intercepts associated with items
!! 7 and 11 resulted in a ML chi-square p-value equal to .004.  This was
!! less than the corrected criterion value of .005, suggesting that
!! at least one of the cross-group equality constraints was misspecified.
!! All tabled p-values less than .005 are marked with an asterisk. In
!! summary, the invariant intercepts hypothesis was rejected
!! for item pairs 7 and 1, 7 and 11, as well as 7 and 20.  The pattern of
!! findings suggested evidence for strong factorial invariance among the
!! following two sets of items: (a) #1, #2, #11, and #20; and (b) #2 and #7.
!! The larger item set was chosen for the partial strong invariance model.
!!
!!         cesd01  cesd02  cesd07  cesd11
!!       --------------------------------
!! cesd02 |  .024        .       .       .
!! cesd07 |  .003*    .029       .       .
!! cesd11 |  .038     .012    .004*      .
!! cesd20 |  .043     .017    .001*   .039
!!
!!
!! Data are from the NHANES 1982-1984 Epidemiological Follow-Up [49,50]
!!----------------------------------------------------------------------

DA  NI=5  NO=248  MA=CM  NG=2

LA
cesd01 cesd02 cesd07 cesd11 cesd20
```

```
CM
0.49587
0.16935    0.57562
0.19368    0.17265    1.29030
0.14082    0.13106    0.31683    0.87469
0.18769    0.12476    0.34361    0.26508    0.66199

ME
0.35081    0.34677    0.80645    0.58871    0.45565


MO NX=5 NK=1 LX=FU,FR TX=FU,FR TD=DI,FR C
             KA=FU,FI PH=SY,FI          !! FIX VALUES OF THE FACTOR
                                        !! MEAN AND VARIANCE

ST 1 PH 1 1                             !! SET THE FACTOR VARIANCE TO EQUAL
                                        !! UNITY

OU  ME=ML


Model for White men ---------------------------------------------------
DA NI=5 NO=2004 MA=CM

LA
cesd01 cesd02 cesd07 cesd11 cesd20

CM
0.49822
0.12634    0.42714
0.20046    0.17796    0.76513
0.18179    0.17049    0.27322    0.82264
0.16644    0.16163    0.31836    0.24146    0.63059

ME
0.32834    0.24002    0.48553    0.61228    0.43713

!! Notes.
!! . LX=IN specifies invariant factor loadings across groups
!! . TX=PS, TD=PS, PH=FU,FR, AND KA=FU,FR specify that item
!!   intercepts and residual variances, as well as the factor
!!   mean and variance are freely estimated for this group
MO NX=5 NK=1 LX=IN TX=PS TD=PS KA=FU,FR PH=FU,FR

!! For C&R-type assessment of partial strong factorial invariance, fit 10
!! separate models. Use the comment command (i.e., '!') to specify each
!! possible pair of the following 5 lines. The following code begins by
!! testing invariance of intercepts associated with items #1 and #2.
 EQ TX(1,1) TX 1  ! Cross-group equality of intercepts for CES-D item01
 EQ TX(1,2) TX 2  ! Cross-group equality of intercepts for CES-D item02
!EQ TX(1,3) TX 3  ! Cross-group equality of intercepts for CES-D item07
!EQ TX(1,4) TX 4  ! Cross-group equality of intercepts for CES-D item11
!EQ TX(1,5) TX 5  ! Cross-group equality of intercepts for CES-D item20

OU ME=ML SO
```