Psychological Bulletin
1989, Vol. 105, No. 3, 456-466

0033-2909/89/$00.75

Testing for the Equivalence of Factor Covariance and Mean Structures: The Issue of Partial Measurement Invariance

Barbara M. Byrne
University of Ottawa, Ottawa, Ontario, Canada

Richard J. Shavelson
University of California, Santa Barbara

Bengt Muthén
University of California, Los Angeles

Addresses issues related to partial measurement invariance using a tutorial approach based on the LISREL confirmatory factor analytic model. Specifically, we demonstrate procedures for (a) using "sensitivity analyses" to establish stable and substantively well-fitting baseline models, (b) determining partially invariant measurement parameters, and (c) testing for the invariance of factor covariance and mean structures, given partial measurement invariance. We also show, explicitly, the transformation of parameters from an all-X to an all-Y model specification, for purposes of testing mean structures. These procedures are illustrated with multidimensional self-concept data from low (n = 248) and high (n = 582) academically tracked high school adolescents.

An important assumption in testing for mean differences is that the measurement (Drasgow & Kanfer, 1985; Labouvie, 1980; Rock, Werts, & Flaugher, 1978) and the structure (Bejar, 1980; Labouvie, 1980; Rock et al., 1978) of the underlying construct are equivalent across groups. One methodological strategy used in testing for this equivalence is the analysis of covariance structures using the LISREL confirmatory factor analytic (CFA) model (Jöreskog, 1971). Although a number of empirical investigations and didactic expositions have used this methodology in testing assumptions of factorial invariance for multiple and single parameters, the analyses have been somewhat incomplete. In particular, researchers have not considered the possibility of partial measurement invariance.

The primary purpose of this article is to demonstrate the application of CFA in testing for, and with, partial measurement invariance. Specifically, we illustrate (a) testing, independently, for the invariance of factor loading (i.e., measurement) parameters, (b) testing for the invariance of factor variance–covariance (i.e., structural) parameters, given partially invariant factor loadings, and (c) testing for the invariance of factor mean structures.[1] Invariance testing across groups, however, assumes well-fitting single-group models; the problem here is to know when to stop fitting the model. A secondary aim of this article, then, is to demonstrate "sensitivity analyses" that can be used to establish stable and substantively meaningful baseline models.

Factorial Invariance

Questions of factorial invariance focus on the correspondence of factors across different groups in the same study, in separate studies, or in subgroups of the same sample (cf. Alwin & Jackson, 1981). The process centers around two issues: measurement invariance and structural invariance. The measurement issue concerns the invariance of regression intercepts, factor loadings (regression slopes), and error/uniqueness variances. The structural issue addresses the invariance of factor mean and factor variance–covariance structures.

Although there are a number of ad hoc methods for comparing factors across independent samples, these procedures were developed primarily for testing the invariance of factors derived from exploratory factor analyses (EFA; see Marsh and Hocevar [1985] and Reynolds and Harding [1983] for reviews). However, Alwin and Jackson (1981) argued that "issues of factorial invariance are not adequately addressed using exploratory factor analysis" (p. 250). A methodologically more sophisticated approach is the CFA method originally proposed by Jöreskog (1971) and now commercially available to researchers through LISREL VI (Jöreskog & Sörbom, 1985) and SPSS^X (SPSS Inc., 1986).[2] (For a discussion of the advantages of CFA over EFA, and details regarding application, see Long [1983], Marsh and Hocevar [1985], and Wolfle [1981].)

LISREL Approach to Testing for Factorial Invariance

As a prerequisite to testing for factorial invariance, it is convenient to consider a baseline model that is estimated separately for each group. This model represents the most parsimonious, yet substantively most meaningful and best fitting, model to the data. Because the χ^2 goodness-of-fit value and its corresponding degrees of freedom are additive, the sum of the χ^2s reflects how

The authors gratefully acknowledge funding support to the first author from the Social Sciences and Humanities Research Council of Canada and computer support from the University of California, Los Angeles. Thanks are also extended to the anonymous reviewers for suggestions that were most helpful in improving the clarity of the paper.

Correspondence concerning this article should be addressed to Barbara M. Byrne, School of Psychology, University of Ottawa, 651 Cumberland, Ottawa, Ontario, Canada K1N 6N5.

[1] In this particular demonstration, we do not test for the invariance of error/uniquenesses. However, the assumption of noninvariant measurement error can also be tested on the basis of partial measurement invariance, as we describe in this article.

[2] Other computer programs available for this procedure include EQS (Bentler, 1985) and COSAN (McDonald, 1978, 1980) for use with interval data and LISCOMP (Muthén, 1987) for use with categorical data.

well the underlying factor structure fits the data across groups. A nonsignificant χ^2 (or a reasonable fit as indicated by some alternate index of fit) is justification that the baseline models fit the observed data.

However, because measuring instruments are often group specific in the way they operate, baseline models are not expected to be identical across groups. For example, whereas the baseline model for one group might include correlated measurement errors, secondary factor loadings,[3] or both, this may not be so for a second group. A priori knowledge of such group differences, as will be illustrated later, is critical to the conduct of invariance-testing procedures. Although the bulk of the literature suggests that the number of factors must be equivalent across groups before further tests of invariance can be conducted, this is only a logical starting point, not a necessary condition; only the comparable parameters within the same factor need to be equated (Werts, Rock, Linn, & Jöreskog, 1976).

Because the estimation of baseline models involves no between-groups constraints, the data may be analyzed separately for each group. In testing for invariance, however, constraints are imposed on particular parameters, and thus the data from all groups must be analyzed simultaneously to obtain efficient estimates (Jöreskog & Sörbom, 1985), with the pattern of fixed and free parameters remaining consistent with that specified in the baseline model for each group.

Tests of factorial invariance, then, can involve both measurement and structural components of a model. In LISREL VI notation, this means that the factor (lambda, Λ), error (theta, Θ), and latent factor variance–covariance (phi, Φ) matrices are of primary interest. If, however, the invariance testing includes factor means, then the regression intercept (nu, ν) and mean (gamma, Γ) vectors are also of interest. More specifically, Λ is a matrix of coefficients regressed from latent factors to observed variables; Θ is the variance–covariance matrix of error/uniquenesses; and ν is a vector of constant intercept terms. These matrices make up the measurement aspect of the model. Φ is the factor variance–covariance matrix, and Γ is a vector of mean estimates. These matrices constitute the structural part of the model. (For a more extensive, yet clear, description of LISREL notation, see Long [1983], Maruyama and McGarvey [1980], and Wolfle [1981].)

Within the Jöreskog tradition, tests of factorial invariance begin with an overall test of the equality of covariance structures across groups (i.e., $H_0: \Sigma_1 = \Sigma_2 = \cdots \Sigma_G$, where G is the number of groups). Failure to reject the null hypothesis is interpreted as evidence of invariance across groups; except for mean structures, the groups can be treated as one. Rejection of this hypothesis, however, leads to testing a series of increasingly restrictive hypotheses in order to identify the source of nonequivalence. These hypotheses relate to the invariance of (a) the factor loading pattern (i.e., $H_0: \Lambda_1 = \Lambda_2 = \cdots \Lambda_G$), (b) the error/uniquenesses (i.e., $H_0: \Theta_1 = \Theta_2 = \cdots \Theta_G$), and (c) the factor variances and covariances (i.e., $H_0: \Phi_1 = \Phi_2 = \cdots \Phi_G$). The tenability of Hypothesis a is a logical prerequisite to the testing of Hypotheses b and c. Recently, however, the importance and rational underlying the omnibus test of equal Σs has been questioned. For a more extensive discussion of this issue, see Byrne (in press).

The procedures for testing the invariance hypotheses are identical to those used in model fitting; that is, a model in which certain parameters are constrained to be equal across groups is compared with a less restrictive model in which these same parameters are free to take on any value. For example, the hypothesis of an invariant pattern of factor loadings (Λ) can be tested by constraining all corresponding lambda parameters to be equal across groups and then comparing this model with one in which the number of factors and the pattern of loadings are invariant but not constrained to be equal (i.e., the summed χ^2s across groups, as mentioned earlier). If the difference in χ^2 ($\Delta\chi^2$) is not significant,[4] the hypothesis of an invariant pattern of loadings is considered tenable. Because the testing for invariant mean structures is more complex, we leave this discussion until later in the article, where a detailed description of the procedure accompanies our example.

In testing for the invariance of factor measurement and variance–covariance structures, modeling with the mean-related parameters does not impose restrictions on the observed variable means; only the analysis of covariance structures is of interest. In testing for the invariance of factor means, on the other hand, the modeling involves restrictions on the observed variable means, and therefore the analysis is based on both the covariance and mean structures.[5] For the purposes of this article, we distinguish between these two phases of invariance testing procedures, both in our summary of the literature and in our application of CFA procedures.

Summary of Literature[6]

Testing for Invariance of Factor Covariance Structures

Factorial invariance has been empirically tested across groups categorized according to socioeconomic status (McGaw & Jöreskog, 1971), race (Wolfle, 1985; Wolfle & Robertshaw, 1983), gender (Marsh, 1985; Marsh, Smith, & Barnes, 1985), educational program (Hanna, 1984), and school type (Lomax, 1985). Additionally, the procedure has been presented didactically with data representing different grade levels (Alwin & Jackson, 1981; Marsh & Hocevar, 1985), reading abilities (Everitt, 1984), race (Alwin & Jackson, 1981; Rock et al., 1978), and socioeconomic strata (Alwin & Jackson, 1981).

Our review of these studies, in addition to examples presented in the LISREL VI manual, revealed two important findings. First, we found no evidence of tests to determine partial measurement invariance. That is, given findings of a noninvariant Λ, Φ, or Θ matrix, no follow-up procedure was implemented or even suggested for pinpointing the source of inequality within the offending matrix. Second, despite findings of a

[3] Secondary loadings are measurement loadings on more than one factor.

[4] The difference in χ^2 ($\Delta\chi^2$) for competing models is itself χ^2 distributed, with degrees of freedom equal to the corresponding difference in degrees of freedom, and indicates whether the reestimated model represents a statistically significant improvement in fit.

[5] For simplicity, we hereinafter refer to these analyses as tests for the invariance of (or differences in) mean structures.

[6] Our review included studies in which the analyses (a) were based on continuous variables, (b) focused primarily on examination of the Λ, Θ, and Φ matrices, and (c) followed procedures as outlined in the LISREL VI manual.

noninvariant Λ matrix, a few researchers continued to test for the invariance of the Φ or Θ matrices.

Testing for Differences in Factor Mean Structures

We found only two published empirical studies that used LISREL CFA procedures to test for differences in latent mean structures (Lomax, 1985; McGaw & Jöreskog, 1971). Several didactic papers, however, have been written on the topic; these papers have included examples that tested across treatment groups (Sörbom, 1982), school settings (Sörbom, 1974), educational programs over time (Hanna & Lei, 1985), and race (Alwin & Jackson, 1981; Everitt, 1984; Rock et al., 1978).

In reviewing these papers, and examples in the LISREL manual, we again found no evidence of tests to determine partial measurement invariance. Furthermore, researchers were consistent in holding entire measurement matrices (Λ, Θ) invariant while testing for latent mean differences. Indeed, we are aware of only one study (Muthén & Christoffersson, 1981) that has tested for latent mean differences using partial measurement invariance. That study, however, focused on the invariance of sets of dichotomous test items and, thus, is only indirectly related to our review.

Our review of methodological procedures for testing the invariance of latent factor and latent mean structures led us to two important conclusions. First, there seems to have been an unfortunate oversight regarding the appropriateness of using the LISREL procedure in testing for partial measurement invariance; the literature is void of any such examples. Consequently, we believe that readers are left with the impression that, given a noninvariant pattern of factor loadings, further testing of invariance and the testing for differences in factor mean scores are unwarranted. This conclusion, however, is unfounded when the model specification includes multiple indicators of a construct and at least one measure (other than the one that is fixed to 1.00 for identification purposes) is invariant (Muthén & Christoffersson, 1981).

Second, consistent with Bentler's (1980) observations almost a decade ago, there is still a paucity of studies that have tested for differences in latent means. It seems evident that an explicit demonstration of the LISREL procedure might be a welcomed addition to the literature. Although, admittedly, there are a number of didactic papers that have outlined this procedure, the presentations assume a fairly high level of statistical sophistication on the part of the reader. For example, we are aware of no paper that explicitly demonstrates how to transform parameters from an all-X to an all-Y model. Although the CFA procedure for testing the factorial invariance of covariance structures is equally valid using either the LISREL all-X or the LISREL all-Y specification, the latter must be used in testing for differences in mean structures (see Everitt, 1984); current didactic papers assume the specification of an all-Y model. We believe that a paper designed to walk the reader through each step of these procedures would make an important contribution to the literature and, perhaps, make the procedure accessible to a wider range of potential users.

We address these limitations by demonstrating procedures for (a) identifying individual noninvariant measurement parameters, (b) testing for the invariance of structural parameters, given partial measurement invariance, (c) respecifying LISREL parameters from an all-X to an all-Y model, and (d) testing for differences in latent factor means, given partial measurement invariance.

Application of LISREL Approach to Tests of Invariance

Data Base

The data derive from a previously published study (Byrne & Shavelson, 1986) and represent multiple self-ratings for each of general self-concept (SC), academic SC, English SC, and mathematics SC for low (n = 248) and high (n = 582) academically tracked students in Grades 11 and 12. These data represent responses to the Self Description Questionnaire (SDQ) III (Marsh & O'Neill, 1984), the Affective Perception Inventory (API; Soares & Soares, 1979), the Self-Esteem Scale (SES; Rosenberg, 1965), and the Self-Concept of Ability Scale (SCA; Brookover, 1962). (For a description of the instruments, and a summary of their psychometric properties, see Byrne and Shavelson [1986]; for a discussion of substantive issues, see Byrne [1988].) Measurements of each SC facet, and a summary of the descriptive statistics for the data, are detailed in the Appendix.

Hypothesized Model

The CFA model in this study hypothesizes a priori that (a) SC responses can be explained by four factors: general SC, academic SC, English SC, and mathematics SC; (b) each subscale measure has a nonzero loading on the SC factor that it is designed to measure (i.e., target loading) and has a zero loading on all other factors (i.e., nontarget loadings); (c) the four SC factors, consistent with the theory (see, e.g., Byrne & Shavelson, 1986), are correlated; and (d) error/uniqueness terms for each of the measures are uncorrelated. Table 1 summarizes the pattern of parameters estimated for the factor loadings (lambda X; Λ_X), factor variance–covariance (phi; Φ), and error variance–covariance (theta delta; Θ_δ) matrices. The λs, ϕs, and θs represent the parameters to be estimated; the zeros and ones were fixed a priori. For purposes of identification, the first of each congeneric set of SC measures was fixed to 1.0 (see, e.g., Long, 1983); each nontarget loading was fixed to 0.0.

Analysis of Data

Analyses were conducted in three stages. First, data for each track were examined separately to establish baseline models. Second, the invariance of SC measurements and structure across track was tested. Finally, latent mean track differences were tested, with equality constraints placed on only those measures known to be invariant across groups.

Although covariance structure analysis has traditionally relied on the χ^2 likelihood ratio test as a criterion for assessing the extent to which a proposed model fits the observed data, its sensitivity to sample size, as well as to various model assumptions (i.e., linearity, multinormality, additivity), is well known (see, e.g., Bentler & Bonett, 1980; Fornell, 1983; Jöreskog, 1982; Marsh & Hocevar, 1985; Muthén & Kaplan, 1985). As an alternative to χ^2, other goodness-of-fit indices have been proposed, albeit their adequacy as criteria of fit has been widely debated in the literature (for a review, see Marsh, Balla, & McDonald, 1988). Overall, researchers have been urged not to

Table 1
Pattern of LISREL Parameters for Model Fitting

$$X = \begin{bmatrix} \text{SDQGSC} \\ \text{APIGSC} \\ \text{SESGSC} \\ \text{SDQASC} \\ \text{SCAASC} \\ \text{SDQESC} \\ \text{APIESC} \\ \text{SCAESC} \\ \text{SDQMSC} \\ \text{APIMSC} \\ \text{SCAMSC} \end{bmatrix} \quad \Lambda_x = \begin{bmatrix} 1 & 0 & 0 & 0 \\ \lambda_{21} & 0 & 0 & 0 \\ \lambda_{31} & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & \lambda_{52} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & \lambda_{73} & 0 \\ 0 & 0 & \lambda_{83} & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & \lambda_{10,4} \\ 0 & 0 & 0 & \lambda_{11,4} \end{bmatrix} \quad \begin{matrix} \xi_1 \\ \text{GSC} \\ \text{ASC} \\ \text{ESC} \\ \text{MSC} \end{matrix}$$

$$\Phi = \begin{bmatrix} \phi_{11} & & & \\ \phi_{21} & \phi_{22} & & \\ \phi_{31} & \phi_{32} & \phi_{33} & \\ \phi_{41} & \phi_{42} & \phi_{43} & \phi_{44} \end{bmatrix}$$

$$\Theta\delta = \begin{bmatrix} \text{SDQGSC} \\ \text{APIGSC} \\ \text{SESGSC} \\ \text{SDQASC} \\ \text{SCAASC} \\ \text{SDQESC} \\ \text{APIESC} \\ \text{SCAESC} \\ \text{SDQMSC} \\ \text{APIMSC} \\ \text{SCAMSC} \end{bmatrix} \begin{bmatrix} \delta_{11} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \delta_{22} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \delta_{33} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \delta_{44} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \delta_{55} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \delta_{66} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \delta_{77} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \delta_{88} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \delta_{99} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \delta_{10,10} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \delta_{11,11} \end{bmatrix}$$

Note. Λ_x = factor loading matrix; Φ = factor variance–covariance matrix; $\Theta\delta$ = error variance–covariance matrix; $\xi_1 - \xi_4$ = SC factors (ξ_1 = general SC; ξ_2 = academic SC; ξ_3 = English SC; ξ_4 = mathematics SC). GSC = general SC: ASC = academic SC; ESC = English SC; MSC = mathematics SC; SDQGSC = Self Description Questionnaire (SDQ)—General Self subscale; APIGSC = Affective Perception Inventory (API)—Self-Concept subscale; SESGSC = Self-Esteem Scale; SDQASC = SDQ Academic SC subscale; SCAASC = Self-Concept of Ability Scale (SCA); SDQESC = SDQ Verbal SC subscale; APIESC = API English Perceptions subscale; SCAESC = SCA Form B (SC of English ability); SDQMSC = SDQ Mathematics SC subscale; APIMSC = API Mathematics Perceptions subscale; SCAMSC = SCA Form C (SC of mathematics ability).

judge model fit solely on the basis of χ^2 values (Bentler & Bonett, 1980; Jöreskog & Sörbom, 1985; Marsh et al., 1988) or on alternative fit indices (Kaplan, 1988; Sobel & Bohrnstedt, 1985); rather, assessments should be based on multiple criteria, including "substantive, theoretical and conceptual considerations" (Jöreskog, 1971, p. 421).

Assessment of overall model fit in the present example was based on statistical as well as practical criteria. Statistical indices of fit included the χ^2 likelihood ratio test, the χ^2/df ratio, and the goodness-of-fit index (GFI) and root-mean-square residual (RMR) provided by LISREL; practical indices included the Bentler and Bonett (1980) normed index (BBI) and the Tucker and Lewis (1973) nonnormed index (TLI).[7] Selection of these indices was based on their widespread use and their usefulness in comparing samples of unequal size (see Marsh et al., 1988). Such popularity notwithstanding, we urge readers to be circumspect in their interpretation of the BBI and TLI because both indices derive from comparison with a null model (see Sobel & Bohrnstedt, 1985). Furthermore, only the TLI has been shown to be relatively independent of sample size (Marsh et al., 1988).

To identify sources of misfit within a specified model, a more detailed evaluation of fit was obtained by inspecting the normalized residuals and modification indices (MIs) provided by LISREL.[8] Additionally, we conducted a sensitivity analysis in order to investigate, under alternative specifications, changes in the estimates of important parameters. Finally, we relied heavily on our knowledge of substantive and theoretical research in SC in making final judgments of the adequacy of a particular model in representing the data.

Model Fit

As is shown in Table 2, the fit of our hypothesized model was poor from a statistical perspective (low track, $\chi^2_{38} = 160.54$; high track, $\chi^2_{38} = 401.09$) and only marginally acceptable from a practical perspective (low track: BBI = .89, TLI = .87; high track: BBI = .92, TLI = .89); this model was therefore rejected.

[7] The GFI indicates the relative amount of variances and covariances jointly explained by the model; it ranges from zero to 1.00, with a value close to 1.00 indicating a good fit. The RMR indicates the average discrepancy between elements in the observed and predicted covariance matrices; it ranges from zero to 1.00, with a value less than .05 being of little practical importance (Sörbom & Jöreskog, 1982). Interpretations based on the BBI and TLI indicate the percentage of covariance explained by the hypothesized model; a value less than .90 usually means that the model can be improved substantially (Bentler & Bonett, 1980).

[8] An MI may be computed for each constrained parameter and indicates the expected decrease in χ^2 if the parameter were to be relaxed; the decrease, however, may actually be higher.

Table 2
Steps in Fitting Baseline Model

Competing models	χ^2	*df*	$\Delta\chi^2$	Δdf	χ^2/df	BBI	TLI
			Low track				
0. Null model	1429.60	55	—	—	25.99	—	—
1. Basic four-factor model with $\theta_i\theta_j = 0$	160.54	38	—	—	4.22	.89	.87
2. $\theta_{10,7}$, free	122.24	37	38.30**	1	3.30	.91	.91
3. $\theta_{10,7}$, θ_{85} free	97.95	36	24.29**	1	2.72	.93	.93
4. $\theta_{10,7}$, θ_{85}, $\theta_{11,5}$ free	71.38	35	26.57**	1	2.04	.95	.96
5. $\theta_{10,7}$, θ_{85}, $\theta_{11,5}$ free, λ_{71} free	54.80	34	16.58**	1	1.61	.96	.98
6. $\theta_{10,7}$, θ_{85}, $\theta_{11,5}$, θ_{96} free, λ_{71} free	49.10	33	5.70*	1	1.49	.97	.98
			High track				
0. Null model	4784.85	55			87.00	—	—
1. Basic four-factor model with $\theta_i\theta_j = 0$	401.09	38	—	—	10.56	.92	.89
2. θ_{85} free	277.67	37	123.42**	1	7.50	.94	.92
3. θ_{85}, $\theta_{11,5}$ free	192.50	36	85.17**	1	5.35	.96	.95
4. θ_{85}, $\theta_{11,5}$, $\theta_{10,7}$ free	153.91	35	38.59**	1	4.40	.97	.96
5. θ_{85}, $\theta_{11,5}$, $\theta_{10,7}$ free, λ_{61} free	126.86	34	27.05**	1	3.73	.97	.97
6. θ_{85}, $\theta_{11,5}$, $\theta_{10,7}$, $\theta_{11,8}$ free, λ_{61} free	105.60	33	21.26**	1	3.20	.98	.97

Note. BBI = Bentler and Bonett (1980) normed index; TLI = Tucker and Lewis (1973) nonnormed index. Dashes indicate not applicable.
* $p < .05$. ** $p < .001$.

Because it was important to establish a well-fitting baseline model for each track separately before testing for factorial invariance, we proceeded in an exploratory fashion to specify a series of alternative models.

Following Sörbom and Jöreskog's (1982) example, we focused on the MIs for each specified model and relaxed, one at a time, only those parameters for which it made sense substantively to do so.[9] For example, previous research with psychological constructs in general (see e.g., Huba, Wingard, & Bentler, 1981; Jöreskog, 1982; Newcomb, Huba, & Bentler, 1986; Sörbom & Jöreskog, 1982; Tanaka & Huba, 1984), and SC constructs in particular (see e.g., Byrne & Schneider, 1988; Byrne & Shavelson, 1986; Byrne & Shavelson, 1987), has demonstrated that in order to obtain a well-fitting model, it is often necessary to allow for correlated errors; such parameter specifications are justified because, typically, they represent nonrandom measurement error due to method effects such as item format associated with subscales of the same measuring instrument. Thus, in fitting our hypothesized four-factor model, we expected, and encountered, similar findings. The highest MIs for each track represented an error covariance between the Mathematics/English Perceptions ($\theta_{10,7}$) subscales of the API (low track, MI = 32.89; high track, MI = 35.49) and between the English/academic SC (θ_{85}; low track, MI = 21.29; high track, MI = 107.71) and mathematics/academic SC ($\theta_{11,5}$; low track, MI = 23.83; high track, MI = 74.70) subscales of the SCA. As is shown in Table 2, each error covariance, when relaxed, resulted in a statistically significant difference in χ^2 for each track. Furthermore, we found a substantial drop in χ^2 when the SDQ Verbal SC (SDQESC; λ_{71}) subscale for the low track ($\Delta\chi^2_1 = 16.58$; MI = 9.11) and the API English Perceptions subscale (APIESC; λ_{61}) for the high track ($\Delta\chi^2_1 = 27.05$; MI = 23.95) were free to load on general SC. These parameters represented secondary loadings, indicating that the two measures of English SC were also tapping perceptions of general SC, a finding that is consistent with previous work in this area.

We considered Model 6 (in Table 2) to be the most plausible baseline model for each track. Although the formal statistical tests of model fit were less than optimal (we address this issue later) for both the low track ($\chi^2_{33} = 49.10$; GFI = .94; RMR = .03) and the high track ($\chi^2_{33} = 105.50$; GFI = .89; RMR = .03), the subjective criteria indicated a theoretically and substantively reasonable representation of the data for both tracks (low track: BBI = .97, TLI = .98; high track: BBI = .98, TLI = .97). This judgment was supported by five additional considerations. First, the data (see the Appendix) approximated a multivariate normal distribution (see Muthén & Kaplan, 1985). Second, all primary factor loading, error variance, and variance–covariance parameters, and their standard errors, were reasonable and statistically significant (see Jöreskog & Sörbom, 1985). Third, the secondary factor loading of APIESC and SDQESC on general SC was substantial for both the low track ($\lambda_{71} = -.33$, $SE = .09$) and the high track ($\lambda_{61} = .17$, $SE = .03$), respectively. Fourth, the assessment instruments, as indicated by the coefficient of determination, performed exceptionally well in measuring the latent SC variables for each track (low track = .998; high track = .996). Finally, the correlated error estimates (a) did not significantly alter the measurement parameter estimates (see Bagozzi, 1983), (b) did not significantly alter the structural parameter estimates (see Fornell, 1983), (c) were significantly different from zero (see Jöreskog, 1982),[10] (d) were considered reasonable because they represented nonrandom error introduced by a particular measurement method (see Gerbing & Anderson, 1984; Sörbom, 1982), and given the size of their loadings, (e) would have an important biasing effect on the other

[9] Although, technically, an MI greater than 3.84 is statistically significant ($\alpha = .05$), we relaxed only parameters that were greater than 5.00 (see Jöreskog & Sörbom, 1985).

[10] The hypothesis that θ_{85}, $\theta_{11,5}$, $\theta_{10,7}$, and $\theta_{11,8}$ for the high track and θ_{85}, $\theta_{11,5}$, $\theta_{10,7}$, and θ_{96} for the low track were equal to zero yielded $\Delta\chi^2_4 = 274.08$ and $\Delta\chi^2_4 = 97.61$ for high and low tracks, respectively.

parameters of interest if constrained to zero (Jöreskog, 1983; see also Alwin & Jackson, 1980).

Given the psychological nature of our sample data, we remained cognizant of two important factors in our determination of baseline models: (a) In the social sciences, hypothesized models must be considered only as approximations to reality rather than as exact statements of truth (Anderson & Gerbing, 1988; Cudeck & Browne, 1983; Jöreskog, 1982); and (b) the sensitivity of χ^2 to sample size is substantially more pronounced for hypothesized target models than for true target models (Marsh et al., 1988). Thus, we directed our efforts toward finding a substantively reasonable approximation to the data and relied more heavily on the practical significance of our findings, and on our own knowledge of the empirical and theoretical work in the area of SC, than on more formal statistical criteria.

Nonetheless, we do not want to leave readers with the impression that statistical criteria are unimportant. Indeed, they provide a vital clue to sources of model misfit. Thus, we now turn to this issue but limit our discussion to the problematic fit of the high-track baseline model only.[11] An examination of the MIs for this model revealed 10 to be greater than 5.00 (highest MI = 15.64); only 2, if relaxed, represented substantively meaningful parameters (θ_{41}, MI = 8.93; $\theta_{10,2}$, MI = 8.23). In an attempt to explain the misfit in these data, we continued fitting the model beyond our selected baseline model. Several additional modifications yielded a statistically better fitting model (χ^2_{26} = 47.96; BBI = .99; TLI = .99) that included three secondary loadings and four correlated error/uniquenesses among subscales of the same instrument.

Such post hoc model fitting has been severely criticized in the literature (see, e.g., Browne, 1982; Cliff, 1983; Fornell, 1983; MacCallum, 1986). However, we argue, as have others (e.g., Huba et al., 1981; Jöreskog, 1982; Tanaka & Huba, 1984), that as long as the researcher is fully cognizant of the exploratory nature of his or her analyses, the process can be substantively meaningful. We prefer to think of the post hoc process as a sensitivity analysis whereby practical, as well as statistical, significances are taken into account. For example, if the estimates of major parameters undergo no appreciable change when minor parameters are added to the model, this is an indication that the initially hypothesized model is empirically robust; the more fitted model therefore represents a minor improvement to an already adequate model, and the additional parameters should be deleted from the model. If, on the other hand, the major parameters undergo substantial alteration, the exclusion of the post hoc parameters may lead to biased estimates (Alwin & Jackson, 1980; Jöreskog, 1983); the minor parameters should therefore be retained in the model.

This suggestion, however, is intended only to serve as a general guide to post hoc model fitting. Clearly, decisions regarding the inclusion or exclusion of parameters must involve the weighing of many additional factors. For example, although the error covariances included in our baseline models affected neither the measurement nor the structural parameters, their absolute values were substantially significant. For this reason, we considered it important to include these parameters in the baseline model for each track.

One method of estimating the practical significance of post hoc parameters is to correlate major parameters (the λs and ϕs) in the baseline model with those in the best-fitting post hoc model. Coefficients close to 1.00 support the stability of the initial model and thus the triviality of the minor parameters in the post hoc model. In contrast, coefficients that are not close to 1.00 (say, <.90) are an indication that the major parameters were adversely affected and thus support the inclusion of the post hoc parameters in the final baseline model.

We subsequently rejected the statistically better fitting model for the high track in favor of the more parsimonious model described in Table 2 (Model 6), on the basis of several considerations. First, the additional secondary factor loadings, although statistically significant, were relatively minor (mean $\hat{\lambda}$ = .06). Second, the additional correlated error/uniquenesses, although statistically significant, were relatively minor (mean $\hat{\theta}$ = .04). Finally, the estimated factor loadings and factor variance–covariances in the baseline model correlated .995 and .992, respectively, with those in the final model, thereby substantiating the stability of the baseline model estimates for the high track. Thus, although the fit for the high track was not statistically optimal, we agree with Cudeck and Browne (1983) that it is sometimes necessary "to sacrifice a little goodness of fit in order to gain interpretability. Clearly, a decision of this nature involves human judgement" (p. 165).

We caution the reader again, however, that this final model was derived from a series of exploratory analyses. Because of capitalization on chance factors, therefore, there is the risk of inflated fit values resulting in possible Type I or Type II errors; caution should be exercised in making substantive interpretations at this point.

Testing Invariance of Factor Structures

The simultaneous estimation of parameters for both tracks was based on the covariance, rather than on the correlation, matrices (see Jöreskog & Sörbom, 1985).[12] And, as was noted earlier, the secondary loading in the Λ matrix, and the correlated errors in the Θ matrix as specified in the baseline model for each group, remained unconstrained throughout the invariance-testing procedures.

Measurement parameters. Because the initial hypothesis of equivalent covariance matrices was rejected, we proceeded, first, to test for the equivalence of SC measurements. These results are presented in Table 3.

The simultaneous four-factor solution for each group yielded a substantively reasonable fit to the data (BBI = .98; TLI = .99). Although these results suggest that for both tracks the data were fairly well described by general SC, academic SC, English SC, and mathematics SC, they do not necessarily imply that the actual factor loadings are the same across track. Thus, the hypothesis of an invariant pattern of factor loadings was tested by constraining all lambda parameters (except λ_{71} and λ_{61})[13] to be

[11] The seemingly better fit for the low track can likely be attributed to the smaller sample size.

[12] Because the model is scale-free, use of the correlation matrix is quite acceptable for single-group analyses; with multiple-group analyses, however, the covariance matrix must be used. The reader is advised that if start values were included in the initial input, they will likely need to be increased in order to make them compatible with covariance rather than correlation values.

[13] Because it was already known that the loadings of λ_{71} and λ_{61} differed across track, the entire lambda matrix was not held invariant.

Table 3
Simultaneous Tests of Invariance for Self-Concept Measurements

Competing models	χ^2	*df*	$\Delta\chi^2$	Δdf
1. Four SC factors invariant	154.60	66	—	—
2. Model 1 with major loadings on each SC factor invariant[a]	180.42	73	25.82***	7
3. Model 1 with major loadings on GSC invariant	154.76	68	0.16	2
4. Model 1 with major loadings on GSC and ASC invariant	162.01	69	7.41	3
5. Model 1 with major loadings on GSC, ASC, and ESC invariant	171.61	71	17.01**	5
6. Model 1 with major loadings on GSC, ASC, and MSC invariant	171.19	71	16.59**	5
7. Model 4 with APIESC invariant	163.30	70	1.29	1
8. Model 4 with SCAESC invariant	166.24	70	4.23*	1
9. Model 4 with APIMSC invariant	169.50	70	7.49**	1
10. Model 4 with SCAMSC invariant	162.06	70	0.05	1

Note. SC = self-concept; GSC = general SC; ASC = academic SC; ESC = English SC; MSC = mathematics SC; APIESC = API English Perceptions subscale; SCAESC = SCA Form B (SC of English ability); APIMSC = API Mathematics Perceptions subscale; SCAMSC = Form C (SC of mathematics ability). Dashes indicate not applicable.
[a] All lambda parameters invariant except λ_{71} and λ_{61}.
* $p < .05$. ** $p < .01$. *** $p < .001$.

equal and then comparing this model (Model 2) with Model 1, in which the number of factors and the pattern of loadings were held invariant across track but not constrained to be equal (see Table 3). The difference in χ^2 was significant ($\Delta\chi^2_7 = 25.82, p < .001$); therefore, the hypothesis of an invariant pattern of factor loadings was untenable.

Because we were interested in pinpointing differences in the measurement parameters between low and high tracks, we proceeded next to test, independently, the invariance of each set of lambda parameters for each SC facet. For example, in examining the measurement of general SC, we held λ_{21} and λ_{31} invariant across track.[14] Given the tenability of this hypothesized model, we next tested the equality of measurements for academic SC by holding λ_{21}, λ_{31}, and λ_{52} invariant. Likewise, we tested the invariance of measurements for both English SC and mathematics SC. In these last two cases, the hypothesis of invariance was rejected. To determine if any of the specific measurements of English and mathematics SC were invariant, we subsequently tested the equality of each of these lambdas individually, while concomitantly holding λ_{21}, λ_{31}, and λ_{52} invariant. The results demonstrated that the measurements of English SC by the SCA (λ_{83}) and of mathematics SC by the API ($\lambda_{10,4}$) were inconsistent across track.[15]

Admittedly, the sequential testing of models in the exploration of partial measurement invariance is problematic. Indeed, given the nonindependence of the tests, it is possible that an alternative series of tests might lead to quite different results. Although we believe that our sequential model-fitting procedures were substantively reasonable, verification must come from cross-validated studies.

Structural parameters. In testing for the equality of SC structure across tracks, we first constrained the entire Φ matrix to be invariant; this hypothesis was found untenable ($\Delta\chi^2_{10} = 47.91, p < .001$). We therefore proceeded to test, independently, the equivalence of each parameter in the Φ matrix; at all times, only those measures known to be consistent in their measurements across track were held invariant (i.e., λ_{21}, λ_{31}, λ_{52}, λ_{73}, $\lambda_{11,4}$). One variance (ϕ_{44}) parameter and two covariance (ϕ_{31}, ϕ_{42}) parameters were found to be noninvariant.[16]

Testing for Differences in Factor Mean Structures

To test for the invariance of mean structures, the use of LISREL required several transformations to our original input (see Table 1). These transformed matrices are presented in Table 4 with the parameter specifications illustrated for each track.

First, the model was restructured into an all-*Y* specification. As such, the factor loading (Λ_X), factor variance–covariance (Φ), and error variance–covariance (Θ_δ) matrices became the Λ_Y, Ψ and Θ_ϵ matrices, respectively; the ξs (the latent factors) were treated as ηs in the LISREL sense. Second, the program was "tricked" into estimating the latent means by the creation of a dummy variable (i.e., an extra variable, DUMMY was added to the variable list, making a total of 12 input variables, not 11). The dummy variable was given a fixed-*X* specification equal to 1.00 (i.e., its value was constrained equal to a value of 1.00). Third, to accommodate the dummy variable, a row of zeros (one for each variable) was added to the last row of the input matrix (which in our study was a correlation matrix), and the value of 1.00 was added to the series of standard deviations (i.e., the standard deviation value representing the dummy variable). Fourth, since the analysis of structured means must be based on the moment, rather than on the covariance matrix, the observed mean values were added to the data input; a value of 1.00 was added for the dummy variable because its value was fixed. Fifth, the Λ and Ψ matrices were modified to accommodate the dummy variable as follows: (a) an extra column of free λs was

[14] The parameter λ_{11} operated as a reference indicator and was therefore fixed to 1.0; likewise, with the first lambda parameter in each set of factor measures.

[15] We also tested for invariance using the more complex model for the high track. With three minor exceptions, the results replicated our findings for the more parsimonious model. One factor covariance parameter (ϕ_{31}; general/English SC) became marginally nonsignificant at the .05 level. Conversely, one factor loading parameter (λ_{52}; SCAASC) and one factor variance parameter (ϕ_{11}; general SC) became marginally significant at the .05 level.

[16] Owing to limitations of space, these results are not reported here but are available from the first author upon request.

Table 4
Pattern of LISREL Parameters for Testing the Invariance of Mean Structures

$$
Y = \begin{bmatrix} \text{SDQGSC} \\ \text{APIGSC} \\ \text{SESGSC} \\ \text{SDQASC} \\ \text{SCAASC} \\ \text{SDQESC} \\ \text{APIESC} \\ \text{SCAESC} \\ \text{SDQMSC} \\ \text{APIMSC} \\ \text{SCAMSC} \end{bmatrix}
$$

$$
\Lambda_Y = \begin{array}{c|ccccc} & \text{Low track} & & & & \\ & \eta_1 & \eta_2 & \eta_3 & \eta_4 & \nu \\ \hline
\text{SDQGSC} & 1 & 0 & 0 & 0 & \lambda_{15} \\
\text{APIGSC} & \lambda_{21} & 0 & 0 & 0 & \lambda_{25} \\
\text{SESGSC} & \lambda_{31} & 0 & 0 & 0 & \lambda_{35} \\
\text{SDQASC} & 0 & 1 & 0 & 0 & \lambda_{45} \\
\text{SCAASC} & 0 & \lambda_{52} & 0 & 0 & \lambda_{55} \\
\text{SDQESC} & 0 & 0 & 1 & 0 & \lambda_{65} \\
\text{APIESC} & \lambda_{71} & 0 & \lambda_{73} & 0 & \lambda_{75} \\
\text{SCAESC} & 0 & 0 & \lambda_{83} & 0 & \lambda_{85} \\
\text{SDQMSC} & 0 & 0 & 0 & 1 & \lambda_{95} \\
\text{APIMSC} & 0 & 0 & 0 & \lambda_{104} & \lambda_{105} \\
\text{SCAMSC} & 0 & 0 & 0 & \lambda_{114} & \lambda_{115}
\end{array}
$$

$$
\begin{array}{c|ccccc} & \text{High track} & & & & \\ & \eta_1 & \eta_2 & \eta_3 & \eta_4 & \nu \\ \hline
\text{SDQGSC} & 1 & 0 & 0 & 0 & \lambda_{15} \\
\text{APIGSC} & \lambda_{21} & 0 & 0 & 0 & \lambda_{25} \\
\text{SESGSC} & \lambda_{31} & 0 & 0 & 0 & \lambda_{35} \\
\text{SDQASC} & 0 & 1 & 0 & 0 & \lambda_{45} \\
\text{SCAASC} & 0 & \lambda_{52} & 0 & 0 & \lambda_{55} \\
\text{SDQESC} & \lambda_{61} & 0 & 1 & 0 & \lambda_{65} \\
\text{APIESC} & 0 & 0 & \lambda_{73} & 0 & \lambda_{75} \\
\text{SCAESC} & 0 & 0 & \lambda_{83} & 0 & \lambda_{85} \\
\text{SDQMSC} & 0 & 0 & 0 & 1 & \lambda_{95} \\
\text{APIMSC} & 0 & 0 & 0 & \lambda_{104} & \lambda_{105} \\
\text{SCAMSC} & 0 & 0 & 0 & \lambda_{114} & \lambda_{115}
\end{array}
$$

$$
\Psi = \begin{array}{c|ccccc} & \eta_1 & \eta_2 & \eta_3 & \eta_4 & \nu \\ \hline
\text{GSC} & \zeta_{11} & & & & \\
\text{ASC} & \zeta_{21} & \zeta_{22} & & & \\
\text{ESC} & \zeta_{31} & \zeta_{32} & \zeta_{33} & & \\
\text{MSC} & \zeta_{41} & \zeta_{42} & \zeta_{43} & \zeta_{44} & \\
\text{DUMMY} & 0 & 0 & 0 & 0 & 0
\end{array}
$$

$$
\begin{array}{c|ccccc} & \eta_1 & \eta_2 & \eta_3 & \eta_4 & \nu \\ \hline
\text{GSC} & \zeta_{11} & & & & \\
\text{ASC} & \zeta_{21} & \zeta_{22} & & & \\
\text{ESC} & \zeta_{31} & \zeta_{32} & \zeta_{33} & & \\
\text{MSC} & \zeta_{41} & \zeta_{42} & \zeta_{43} & \zeta_{44} & \\
\text{DUMMY} & 0 & 0 & 0 & 0 & 0
\end{array}
$$

$$
\theta_\epsilon = \begin{array}{c|ccccccccccc} & \text{SDQGSC} & \text{APIGSC} & \text{SESGSC} & \text{SDQASC} & \text{SCAASC} & \text{SDQESC} & \text{APIESC} & \text{SCAESC} & \text{SDQMSC} & \text{APIMSC} & \text{SCAMSC} \\ \hline
\text{SDQGSC} & \epsilon_{11} & & & & & & & & & & \\
\text{APIGSC} & 0 & \epsilon_{22} & & & & & & & & & \\
\text{SESGSC} & 0 & 0 & \epsilon_{33} & & & & & & & & \\
\text{SDQASC} & 0 & 0 & 0 & \epsilon_{44} & & & & & & & \\
\text{SCAASC} & 0 & 0 & 0 & 0 & \epsilon_{55} & & & & & & \\
\text{SDQESC} & 0 & 0 & 0 & 0 & 0 & \epsilon_{66} & & & & & \\
\text{APIESC} & 0 & 0 & 0 & 0 & 0 & 0 & \epsilon_{77} & & & & \\
\text{SCAESC} & 0 & 0 & 0 & 0 & \epsilon_{85} & 0 & 0 & \epsilon_{88} & & & \\
\text{SDQMSC} & 0 & 0 & 0 & 0 & 0 & \epsilon_{96} & 0 & 0 & \epsilon_{99} & & \\
\text{APIMSC} & 0 & 0 & 0 & 0 & 0 & 0 & \epsilon_{107} & 0 & 0 & \epsilon_{1010} & \\
\text{SCAMSC} & 0 & 0 & 0 & 0 & \epsilon_{115} & 0 & 0 & 0 & 0 & 0 & \epsilon_{1111}
\end{array}
$$

$$
\begin{array}{c|ccccccccccc} & \text{SDQGSC} & \text{APIGSC} & \text{SESGSC} & \text{SDQASC} & \text{SCAASC} & \text{SDQESC} & \text{APIESC} & \text{SCAESC} & \text{SDQMSC} & \text{APIMSC} & \text{SCAMSC} \\ \hline
\text{SDQGSC} & \epsilon_{11} & & & & & & & & & & \\
\text{APIGSC} & 0 & \epsilon_{22} & & & & & & & & & \\
\text{SESGSC} & 0 & 0 & \epsilon_{33} & & & & & & & & \\
\text{SDQASC} & 0 & 0 & 0 & \epsilon_{44} & & & & & & & \\
\text{SCAASC} & 0 & 0 & 0 & 0 & \epsilon_{55} & & & & & & \\
\text{SDQESC} & 0 & 0 & 0 & 0 & 0 & \epsilon_{66} & & & & & \\
\text{APIESC} & 0 & 0 & 0 & 0 & 0 & 0 & \epsilon_{77} & & & & \\
\text{SCAESC} & 0 & 0 & 0 & 0 & \epsilon_{85} & 0 & 0 & \epsilon_{88} & & & \\
\text{SDQMSC} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \epsilon_{99} & & \\
\text{APIMSC} & 0 & 0 & 0 & 0 & 0 & 0 & \epsilon_{107} & 0 & 0 & \epsilon_{1010} & \\
\text{SCAMSC} & 0 & 0 & 0 & 0 & \epsilon_{115} & 0 & 0 & \epsilon_{118} & 0 & 0 & \epsilon_{1111}
\end{array}
$$

$$
\Gamma = \begin{array}{c|c} & \\ \hline
\text{GSC} & 0 \\
\text{ASC} & 0 \\
\text{ESC} & 0 \\
\text{MSC} & 0 \\
\text{DUMMY} & 1
\end{array}
$$

$$
\begin{array}{c|c} & \\ \hline
\text{GSC} & \gamma_{11} \\
\text{ASC} & \gamma_{21} \\
\text{ESC} & \gamma_{31} \\
\text{MSC} & \gamma_{41} \\
\text{DUMMY} & 1_{22}
\end{array}
$$

Note. Λ_Y = factor loading matrix; Ψ = factor variance–covariance matrix; θ_ϵ = error variance–covariance matrix; Γ = mean estimate vector; Y = observed measures of self-concept (SC); $\eta_1 - \eta_4$ = SC factors (η_1 = general SC; η_2 = academic SC; η_3 = English SC; η_4 = mathematics SC); ν = mean intercepts; GSC = general SC; ASC = academic SC; ESC = English SC; MSC = mathematics SC; DUMMY = extra variable; SDQGSC = Self Description Questionnaire (SDQ) General Self subscale; APIGSC = Affective Perception Inventory (API)—Self-Concept subscale; SESGSC = Self-Esteem Scale; SDQASC = SDQ Academic SC subscale; SCAASC = Self-Concept of Ability Scale (SCA); SDQESC = SDQ English SC subscale; APIESC = API English Perceptions subscale; SCAESC = SCA Form B (SC of English ability); SDQMSC = SDQ Mathematics SC subscale; APIMSC = Mathematics Perceptions subscale; SCAMSC = SCA Form C (SC of mathematics ability).

added to the Λ matrix, which represented the measurement intercepts; and (b) an extra row of zeros was added to the Ψ matrix, and ζ_{55} was fixed to zero.[17] Finally, the latent mean values were estimated in the gamma (Γ) matrix. Whereas γ_{11} to γ_{41} were fixed to zero for the low track, these parameters were freely estimated for the high track; γ_{51} was fixed to 1.0 for both tracks.

Because the origins of the measurements and the means of the latent variables cannot be identified simultaneously, absolute mean estimates are not possible. However, when the parameter specifications, as were described earlier, are imposed, latent mean differences between groups can be estimated; one group is used as the reference group, and as such, its latent mean parameters are fixed to 0.0. In this case, the low track served as the reference group; mean parameters for the high track were freely estimated. Comparison of the groups, then, was based on the difference from zero. Statistical significance was based on T values (mean estimates divided by their standard error estimates).

We emphasize, again, that only the factor loading parameters known to be consistent in their SC measurements across track were held invariant. In particular, the reader should note that because λ_{71} and λ_{61} were freely estimated for the low track and the high track, respectively, the intercept terms for these parameters (λ_{75}, λ_{65}) were also free to vary for each track.

The parameter estimates and standard errors are presented in Table 5. Examination of the gamma estimates ($\gamma_{11} - \gamma_{41}$) revealed statistically significant mean track differences in academic, English, and mathematics SCs, with positive values indicating higher scores for the high track. The largest differences between tracks were in academic SC (γ_{21}), followed by mathematics SC (γ_{41}) and English SC (γ_{31}), respectively. Mean track

[17] The LISREL program will print the message, "PSI is not positive definite." This can be ignored because it is a function of ζ_{55} being fixed to 0.0.

Table 5
Maximum Likelihood Estimates and Standard Errors for Self-Concept Facets

	Low track		Across-track equivalencies		High track	
Parameter	Estimate	*SE*	Estimate	*SE*	Estimate	*SE*
ν_1 (λ_{15})			75.71	0.81		
ν_2 (λ_{25})			76.69	0.47		
ν_3 (λ_{35})			31.34	0.29		
ν_4 (λ_{45})			47.55	0.67		
ν_5 (λ_{55})			25.20	0.28		
ν_6 (λ_{65})	55.07	0.61			52.35	0.69
ν_7 (λ_{75})	58.03	0.65			54.36	0.88
ν_8 (λ_{85})			25.61	0.31		
ν_9 (λ_{95})			41.72	0.81		
ν_{10} ($\lambda_{10,5}$)			42.17	0.58		
ν_{11} ($\lambda_{11,5}$)			23.05	0.35		
λ_{21}			16.00	0.66		
λ_{31}			10.69	0.35		
λ_{52}			13.72	0.56		
λ_{73}			42.71	1.94		
λ_{83}	13.71	1.31			18.53	0.94
$\lambda_{10,4}$	23.98	1.21			20.26	0.43
$\lambda_{11,4}$			12.83	0.31		
λ_{61}					3.56	0.69
λ_{71}	−6.98	1.66				
$\theta\epsilon_{11}$	40.03	6.81			44.72	4.86
$\theta\epsilon_{22}$	42.06	4.29			40.33	2.69
$\theta\epsilon_{33}$	6.38	0.94			4.25	0.57
$\theta\epsilon_{44}$	77.80	8.72			52.95	4.16
$\theta\epsilon_{55}$	8.88	1.34			8.05	0.78
$\theta\epsilon_{66}$	42.53	4.76			37.12	2.73
$\theta\epsilon_{77}$	28.16	6.30			17.67	3.38
$\theta\epsilon_{88}$	14.43	1.48			14.34	1.05
$\theta\epsilon_{99}$	41.86	5.72			25.91	3.22
$\theta\epsilon_{10,10}$	22.79	3.65			16.56	1.65
$\theta\epsilon_{11,11}$	10.92	1.25			14.30	1.00
$\theta\epsilon_{85}$	4.67	0.96			6.12	0.66
$\theta\epsilon_{10,7}$	17.60	3.17			7.65	1.41
$\theta\epsilon_{11,5}$	3.70	0.85			5.69	0.65
$\theta\epsilon_{96}$	−8.37	3.63				
$\theta\epsilon_{11,8}$					3.20	0.72
γ_{11} (GSC)	0.0				0.01	0.03
γ_{21} (ASC)	0.0				0.36	0.03
γ_{31} (ESC)	0.0				0.17	0.02
γ_{41} (MSC)	0.0				0.25	0.03
ζ_{11}	0.15	0.12			0.19	0.01
ζ_{22}	0.07	0.01			0.09	0.01
ζ_{33}	0.05	0.01			0.06	0.01
ζ_{44}	0.15	0.02			0.29	0.02

Note. $\chi^2_{(76)} = 201.82$. GSC = general self-concept (SC); ASC = academic SC: ESC = English SC; MSC = mathematics SC.

differences in general SC (γ_{11}) were negligible and not statistically significant.

The results demonstrate that, overall, the test for invariant SCs across track based on mean and covariance structures was statistically more powerful than tests based on covariance structures alone. Whereas tests of invariance based on the latter found academic track differences in mathematics SC (ϕ_{44}) only, this was not so in the analysis that also included mean structures. Significant differences were also found in academic and English SCs.

Conclusion

Using data based on SC responses for low- and high-track high school students, we demonstrated procedures for (a) establishing a substantively well-fitting baseline model for each group; (b) conducting sensitivity analyses to assess the stability of a baseline model; (c) determining partial measurement invariance by testing parameters, independently, given findings of noninvariance at the matrix level; and (d) testing for factorial invariance and differences in mean structures, given partially invariant measuring instruments. Throughout the article, we emphasized the exploratory nature of our analyses and noted the limitations of interpretations based on the results.

Post hoc analyses with confirmatory covariance structure models are, indeed, problematic; with multiple model respecifications, probability values become meaningless. At this point in time, however, there is simply no direct way to adjust for the probability of either Type I or Type II errors arising from

capitalization on chance factors (see also Cliff, 1983). This represents a serious limitation in the analysis of covariance and mean structures because, realistically, most psychological research is likely to require the specification of alternative models in order to attain one that is well fitting (see, e.g., Anderson & Gerbing, 1988; MacCallum, 1986). Thus, practitioners of the LISREL methodology must await the research efforts of statisticians in resolving this important psychometric obstacle.

In the meantime, one approach to the problem is to employ a cross-validation strategy using an independent sample (Anderson & Gerbing, 1988; Bentler, 1980; Browne, 1982; Cliff, 1983; Cudeck & Browne, 1983; Long, 1983; MacCallum, 1986). In lieu of collecting new data, one can randomly subdivide a large sample into two. The researcher uses exploratory procedures with the first subsample to determine a well-fitting model; hypotheses related to this model are then tested, statistically, using confirmatory procedures on data from the second subsample. In this way, the model is not influenced by the data, and thus the hypotheses can be tested legitimately within a confirmatory framework (Cliff, 1983).

Cross-validation, however, is not a panacea; it requires judicial implementation. The procedure is most effective with minimal model modifications; the relaxation of many parameters is likely to yield an unsuccessful cross-validation. The major disadvantage of cross-validation, of course, is reduced sample size. However, as others have noted (Cliff, 1983; Cudeck & Browne, 1983), the benefits derived in estimate stability and interpretability far outweigh this limitation. Alternatively, Cudeck and Browne (1983) have outlined cross-validation procedures for use with small samples, in which case sample splitting may be statistically inappropriate.

In this article, we have encouraged researchers to gather maximal information regarding individual model parameters, and we have provided some technical details on how to do so. However, as with other statistical procedures, this information comes at a price: the risk of capitalization on chance factors. Thus, we emphasize the importance of exercising sound judgment in the implementation of these procedures; one should not relax constrained parameters unless it makes sense substantively to do so. Only a solid theoretical and substantive knowledge of one's subject area can guide this investigative process. Cross-validation procedures can then be used to test for the validity of these results.

References

Alwin, D. F., & Jackson, D. J. (1980). Measurement models for response errors in surveys: Issues and applications. In K. F. Schuessler (Ed.), *Sociological methodology* (pp. 68–119). San Francisco: Jossey-Bass.

Alwin, D. F., & Jackson, D. J. (1981). Applications of simultaneous factor analysis to issues of factorial invariance. In D. D. Jackson & E. P. Borgatta (Eds.), *Factor analysis and measurement in sociological research: A multidimensional perspective* (pp. 249–280). Beverly Hills, CA: Sage.

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin, 103,* 411–423.

Bagozzi, R. P. (1983). Issues in the application of covariance structure analysis: A further comment. *Journal of Consumer Research, 9,* 449–450.

Bejar, I. I. (1980). Biased assessment of program impact due to psychometric artifacts. *Psychological Bulletin, 87,* 513–524.

Bentler, P. M. (1980). Multivariate analysis with latent variables: Causal modeling. *Annual Review of Psychology, 31,* 419–456.

Bentler, P. M. (1985). *Theory and implementation of EQS: A structural equations program.* Los Angeles, CA: BMDP Statistical Software.

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

Brookover, W. B. (1962). *Self-concept of Ability Scale.* East Lansing, MI: Educational Publication Services.

Browne, M. W. (1982). Covariance structures. In D. M. Hawkins (Ed.), *Topics in applied multivariate analysis* (pp. 72–141). New York: Cambridge University Press.

Byrne, B. M. (1988). Adolescent self-concept, ability grouping, and social comparison: Reexamining academic track differences in high school. *Youth and Society, 20,* 46–67.

Byrne, B. M. (in press). *A primer of LISREL: Basic applications and programming for confirmatory factor analytic models.* New York: Springer-Verlag.

Byrne, B. M., & Schneider, B. H. (1988). Perceived Competence Scale for Children: Testing for factorial validity and invariance across age and ability. *Applied Measurement in Education, 1,* 171–187.

Byrne, B. M., & Shavelson, R. J. (1986). On the structure of adolescent self-concept. *Journal of Educational Psychology, 78,* 474–481.

Byrne, B. M., & Shavelson, R. J. (1987). Adolescent self-concept: Testing the assumption of equivalent structure across gender. *American Educational Research Journal, 24,* 365–385.

Cliff, N. (1983). Some cautions of causal modeling methods. *Multivariate Behavioral Research, 18,* 115–126.

Cudeck, R., & Browne, M. W. (1983). Cross-validation of covariance structures. *Multivariate Behavioral Research, 18,* 147–167.

Drasgow, F., & Kanfer, R. (1985). Equivalence of psychological measurement in heterogeneous populations. *Journal of Applied Psychology, 70,* 662–680.

Everitt, B. S. (1984). *An introduction to latent variable models.* New York: Chapman and Hall.

Fornell, C. (1983). Issues in the application of covariance structure analysis: A comment. *Journal of Consumer Research, 9,* 443–448.

Gerbing, D. W., & Anderson, J. C. (1984). On the meaning of within-factor correlated measurement errors. *Journal of Consumer Research, 11,* 572–580.

Hanna, G. (1984). The use of a factor-analytic model for assessing the validity of group comparisons. *Journal of Educational Measurement, 21,* 191–199.

Hanna, G., & Lei, H. (1985). A longitudinal analysis using the LISREL-model with structured means. *Journal of Educational Statistics, 10,* 161–169.

Huba, G. J., Wingard, J. A., & Bentler, P. M. (1981). A comparison of two latent-variable causal models for adolescent drug use. *Journal of Personality and Social Psychology, 40,* 180–193.

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

Jöreskog, K. G. (1982). Analysis of covariance structures. In C. Fornell (Ed.), *A second generation of multivariate analysis: Vol. 1. Methods* (pp. 200–242). New York: Praeger.

Jöreskog, K. G. (1983, August). *UK LISREL workshop* [Workshop sponsored by Centre for Educational Sociology]. University of Edinburgh, Scotland.

Jöreskog, K. G., & Sörbom, D. (1985). *LISREL VI: Analysis of linear structural relationships by the method of maximum likelihood.* Mooresville, IN: Scientific Software.

Kaplan, D. (1988). The impact of specification error on the estimation, testing, and improvement of structural equation models. *Multivariate Behavioral Research, 23,* 69–86.

Labouvie, E. W. (1980). Identity versus equivalence of psychological measures and constructs. In L. W. Poon (Ed.), *Aging in the 1980's* (pp. 493–502). Washington, DC: American Psychological Association.

Lomax, R. G. (1985). A structural model of public and private schools. *Journal of Experimental Education, 53,* 216–226.

Long, J. S. (1983). *Confirmatory factor analysis.* Beverly Hills, CA: Sage.

MacCallum, R. (1986). Specification searches in covariance structure modeling. *Psychological Bulletin, 100,* 107–120.

Marsh, H. W. (1985). The structure of masculinity/femininity: An application of confirmatory factor analysis to higher-order factor structures and factorial invariance. *Multivariate Behavioral Research, 20,* 427–449.

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

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.

Marsh, H. W., & O'Neill, R. (1984). Self Description Questionnaire III: The construct validity of multidimensional self-concept ratings by late adolescents. *Journal of Educational Measurement, 21,* 153–174.

Marsh, H. W., Smith, I. D., & Barnes, J. (1985). Multidimensional self-concepts: Relations with sex and academic achievement. *Journal of Educational Psychology, 77,* 581–596.

Maruyama, G., & McGarvey, B. (1980). Evaluating causal models. An application of maximum-likelihood analysis of structural equation models. *Psychological Bulletin, 87,* 502–512.

McDonald, R. P. (1978). A simple comprehensive model for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology, 31,* 59–72.

McDonald, R. P. (1980). A simple comprehensive model for the analysis of covariance structures: Some remarks on applications. *British Journal of Mathematical and Statistical Psychology, 33,* 161–183.

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, 24,* 154–168.

Muthén, B. (1987). *LISCOMP: Analysis of linear structural relations with a comprehensive measurement model.* Mooresville, IN: Scientific Software.

Muthén, B., & Christoffersson, A. (1981). Simultaneous factor analysis of dichotomous variables in several groups. *Psychometrika, 46,* 407–419.

Muthén, B., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology, 38,* 171–189.

Newcomb, M. D., Huba, G. T., & Bentler, P. M. (1986). Determinants of sexual and dating behaviors among adolescents. *Journal of Personality and Social Psychology, 50,* 428–438.

Reynolds, C. R., & Harding, R. E. (1983). Outcome in two large sample studies of factorial similarity under six methods of comparison. *Educational and Psychological Measurement, 43,* 723–728.

Rock, D. A., Werts, C. E., & Flaugher, R. L. (1978). The use of analysis of covariance structures for comparing the psychometric properties of multiple variables across populations. *Multivariate Behavioral Research, 13,* 403–418.

Rosenberg, M. (1965). *Society and the adolescent self-image.* Princeton, NJ: Princeton University Press.

Soares, A. T., & Soares, L. M. (1979). *The Affective Perception Inventory—Advanced level.* Trumbell, CN: ALSO.

Sobel, M. E., & Bohrnstedt, G. W. (1985). Use of null models in evaluating the fit of covariance structure models. In N. B. Tuma (Ed.), *Sociological methodology* (pp. 152–178). San Francisco: Jossey-Bass.

Sörbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. *British Journal of Mathematical and Statistical Psychology, 27,* 229–239.

Sörbom, D. (1982). Structural equation models with structured means. In K. G. Jöreskog & H. Wold (Eds.), *Systems under direct observation* (pp. 183–195). Amsterdam: North-Holland.

Sörbom, D., & Jöreskog, K. G. (1982). The use of structural equation models in evaluation research. In C. Fornell (Ed.), *A second generation of multivariate analysis: Vol. 2. Measurement and evaluation* (pp. 381–418). New York: Praeger.

SPSS Inc. (1986). *SPSS^x user's guide.* New York: McGraw-Hill.

Tanaka, J. S., & Huba, G. J. (1984). Confirmatory hierarchical factor analyses of psychological distress measures. *Journal of Personality and Social Psychology, 46,* 621–635.

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

Werts, C. E., Rock, D. A., Linn, R. L., & Jöreskog, K. G. (1976). Comparison of correlations, variances, covariances, and regression weights with or without measurement error. *Psychological Bulletin, 83,* 1007–1013.

Wolfle, L. M. (1981, April). *Causal models with unmeasured variables: An introduction to LISREL.* Paper presented at the annual meeting of the American Educational Research Association, Los Angeles, CA.

Wolfle, L. M. (1985). Postsecondary educational attainment among Whites and Blacks. *American Educational Research Journal, 22,* 501–525.

Wolfle, L. M., & Robertshaw, D. (1983). Racial differences in measurement error in educational achievement models. *Journal of Educational Measurement, 20,* 39–49.

Appendix

Summary of Self-Concept Measurements and Descriptive Summary of Data

Self-concept factors are general self-concept, academic self-concept, English self-concept, and mathematics self-concept. Measures of self-concept factors are as follows: general self-concept—SDQ General Self subscale, API Self-Concept subscale, SES; academic self-concept[A1]—SDQ Academic Self-Concept subscale, SCA Form A; English self-concept—SDQ English Self-Concept subscale, API English Perceptions subscale, SCA Form B; mathematics self-concept—SDQ Mathematics Self-Concept subscale, API Mathematics Perceptions subscale, SCA Form C.

On the basis of listwise deletion of missing cases, the data were considered to approximate a normal distribution. Skewness ranged from −1.19 to .19 (M = −.27) for the low track and from −1.26 to .10 (M = −.50) for the high track; kurtosis ranged from −.53 to 1.60 (M = .23) for the low track and from −.92 to 1.83 (M = .27) for the high track.

[A1] Although the API Student Self subscale was originally intended as one measure of academic SC, a factor analysis in an earlier study (Byrne & Shavelson, 1986) showed this subscale to be problematic; only 10 of the 25 items loaded greater than .25 on the academic SC factor. We therefore deleted it as a measure of academic SC in this study.

Received February 23, 1987
Revision received January 21, 1988
Accepted August 1, 1988 ■