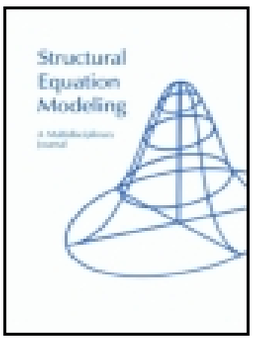

Routledge
Taylor & Francis Group

Structural Equation Modeling: A Multidisciplinary Journal

ISSN: 1070-5511 (Print) 1532-8007 (Online) Journal homepage: http://www.tandfonline.com/loi/hsem20

Longitudinal stability of latent means and individual differences: A unified approach

Herbert W. Marsh & David Grayson

To cite this article: Herbert W. Marsh & David Grayson (1994) Longitudinal stability of latent means and individual differences: A unified approach, Structural Equation Modeling: A Multidisciplinary Journal, 1:4, 317-359, DOI: 10.1080/10705519409539984

To link to this article: http://dx.doi.org/10.1080/10705519409539984

STRUCTURAL EQUATION MODELING, 1(2), 317–359

Longitudinal Stability of Latent Means and Individual Differences: A Unified Approach

Herbert W. Marsh
University of Western Sydney, Australia

David Grayson
University of Sydney, Australia

We examined the stability of responses to a multi-item self-esteem scale collected on five occasions over an 8-year period. A wide variety of approaches were critically examined that considered the stability of means, individual differences (i.e., test–retest correlations), and factor structures using traditional approaches (e.g., ANOVA and correlations) and structural equation models. Structural equation models based on multiple indicators provided a unified analytic approach for evaluating different aspects of stability and offered important advantages over traditional approaches. We describe a hierarchy of invariances and the nature of interpretations that are justified by different patterns of factor structure invariance associated with each level. We conclude that the assumptions underlying the typical repeated-measures ANOVA approach to testing mean differences in longitudinal data are far more restrictive, less easily tested, and less likely to be met than those in the structural equation modeling approach advocated here, and that the use of ANOVA for this purpose requires a huge leap of faith that can rarely be justified on logical or empirical grounds.

This investigation is a methodological demonstration of different approaches to evaluating stability in longitudinal, multiwave panel studies. We demonstrate these approaches in relation to substantively important issues in

Requests for reprints should be sent to Herbert W. Marsh, Faculty of Education, University of Western Sydney–Macarthur, P.O. Box 555, Campbelltown, New South Wales 2560, Australia.

the study of self-concept, but emphasize that these techniques are relevant to a wide variety of multiwave panel studies and the longitudinal analyses of individual differences. In this respect, the study has important methodological and substantive implications that are mutually reinforcing.

STABILITY OF SELF-CONCEPT: THE SUBSTANTIVE ISSUE

Self-concept theorists are concerned with the stability of self-concept. Rosenberg (1985) emphasized the need for an individual to have a reasonably stable self-concept because "the self-concept is his most fundamental frame of reference; without a firm clear picture of what one is like, the individual is virtually immobilized" (p. 220). Rosenberg also noted similar conclusions by other theorists. Markus and Kunda (1986) argued that "the most pervasive and least ambiguous finding to emerge from the recent surge of research on self-concept is that individuals seek out consistency and stability and actively resist any information that challenges their prevailing view of themselves" (p. 858). There has been particular interest in the stability of self-concept during the adolescent period, which has been characterized (Dusek & Flaherty, 1981) as a time of storm and stress. In contrast to this characterization, there is growing evidence that self-concept is relatively stable during adolescence and that growth is gradual and continuous (e.g., Dusek & Flaherty, 1981; Marsh, 1989; O'Malley & Bachman, 1983). These are primarily concerns with the stability of individual differences in self-concept as inferred from test–retest correlations across two or more occasions.

Self-concept researchers are also concerned with developmental changes in self-concept over time. Here interest focuses on changes in the mean levels of self-concept as a function of age. Marsh (1989), for example, reviewed previous research and presented analyses of responses by over 12,000 individuals varying in age from early childhood to young adulthood. Across many different studies and various components of self-concept, a reasonably consistent picture arose. For responses by preadolescents there was a linear decline in self-concept with age. During early adolescence there was a reasonably consistent quadratic effect; the decline in self-concept continued through Grades 8 and 9 (about the age of puberty), and then began to increase. For responses by late adolescents there was a reasonably consistent linear increase in self-concept that continued through at least early adulthood. Whereas most of this research consisted of cross-sectional studies in which changes in self-concept were inferred from the results of between-group analyses, a few of the studies were true longitudinal studies that used repeated-measures analyses.

Although the self-esteem data provides an intuitively appealing vehicle for considering these methodological issues, it is important to emphasize that

the issues considered here are fundamental to the interpretation of results for all repeated-measures data. The basic issues addressed in self-concept research—the stability of individual differences, the continuity of growth, and changes in mean levels—are relevant to all individual difference variables. Whereas multiwave, longitudinal data provide an appropriate basis for evaluating these issues, researchers have often used weak or inappropriate statistical techniques. In particular, researchers have typically considered issues related to the stability of individual differences separately from evaluations of changes in mean levels of self-concept. Here we outline a unified approach that evaluates both questions within a single analytic framework. In illustrating this approach, we demonstrate potentially serious limitations in the typical repeated-measures analysis used to evaluate mean differences in longitudinal data.

STATISTICAL ANALYSES OF MULTIWAVE LONGITUDINAL DATA: THE METHODOLOGICAL ISSUE

For purposes of this methodological demonstration, we consider the stability of self-esteem inferred from responses to a multiitem measure collected on five occasions over an 8-year period. For these data we ask whether there are systematic differences in the mean level of self-esteem over time and how highly correlated self-esteem is from one occasion to the next. We compare traditional approaches based on raw scale scores for inferring mean differences (e.g., repeated-measures analysis of variance; ANOVA) and relations among raw scores (e.g., test–retest correlations) with structural equation models in which latent constructs are inferred from multiple indicators.

There are many approaches to the study of stability and change (Bock, 1985; Collins & Horn, 1991; Goldstein, 1979; McArdle, 1988; Meredith, 1991; Plewis, 1985; Rogosa, 1979; Willett, 1988). The aim of our presentation is not to provide a new, unique technique or statistical model of change. Rather, we attempt to provide an applied perspective on how issues fundamental to the analysis of longitudinal data and the interpretation of change can be addressed in a structural equation modeling (SEM) framework that incorporates multiple indicators of each latent construct.

The two most common issues in the evaluation of longitudinal data refer to the stability of means over time (mean stability) and to the stability of individual differences over time (covariance stability). Although many studies attempt to infer mean stability or growth on the basis of cross-sectional data, the advantages of longitudinal data for this purpose are well known, and the assessment of covariance stability studies requires longitudinal data. Evaluations of stability typically focus on either the stability of means over time or the stability of individual differences as inferred by covariance

stability, but not both. The purpose of this investigation is to evaluate approaches to inferring both types of stability and demonstrate the application of structural equation models that incorporate both approaches into a single analytic framework. Recent advances in the application of structural equation models (e.g., Byrne, Shavelson, & Muthen, 1989; Jöreskog & Sörbom, 1988; Marsh & Grayson, 1990) allow researchers to compare the psychometric properties of the same measures across multiple groups, to compare latent means for the different groups, and to test the appropriateness of interpretations of these data. Here we demonstrate generalizations of the multigroup procedures to compare psychometric properties of the same instrument across multiple occasions within a single group, to compare means on latent constructs for the multiple occasions, and to test the appropriateness of interpretations of these data.

METHOD

Sample and Procedures

Data came from the large, nationally representative Youth in Transition study of all 10th grade boys in public high schools in 1966 (Bachman, 1970, 1975). A two-stage sampling scheme was used in which a random sample of 87 public high schools was selected and then approximately 25 students were randomly selected from each school. The commercially available data are from Waves 1 (early 10th grade; $N = 2213$), 2 (late 11th grade; $N = 1886$), 3 (late 12th grade; $N = 1799$), 4 (1 year after normal high school graduation; $N = 1620$), and 5 (5 years after normal high school graduation; $N = 1,608$). For each of the five data waves, self-esteem was inferred from responses to the same 10 items (see Table 1). In the commercially available data (Bachman,

TABLE 1
Wording of the Items on the Self-Esteem Scale

I feel that I'm a person of worth, at least on an equal plane with others.
I feel that I have a number of good qualities.
I am able to do things as well as most people.
I feel that I do not have much to be proud of.[a]
I take a positive attitude toward myself.
Sometimes I think I am no good at all.[a]
I am a useful guy to have around.
I feel that I can't do anything right.[a]
When I do a job, I do it well.
I feel that my life is not very useful.[a]

Note. All responses on a 5-point scale ranging from *never true, seldom true, sometimes true, often true,* and *almost always true.*

[a]Reverse scored.

1975), scale scores were defined as the mean of nonmissing responses so long as students responded to at least 8 of the 10 items for any given wave, and this strategy was used here. For our purposes, responses were considered for the 1,341 students who responded to at least 8 of 10 items on all five occasions. For each wave, responses to the same 10 items were used to compute 5 item-pairs (item parcels or subscales) by computing the average response to the first 2 items, the second 2 items, and so forth. Across all five waves there were 25 measured variables (item-pairs; see Appendix A). Because of the two-stage cluster sampling scheme used in the original data collection, standard errors based on the assumption of simple random sampling are biased. Bachman (1970) systematically evaluated appropriate design effects to compensate for this bias. Based on this research, Bachman and O'Malley (1986) suggested that an N of 1,000 should be used for purposes of testing statistical significance for a study based on a slightly larger number of students than considered here, and we adopted this value for our purposes. It should be emphasized that this reduction in the nominal sample size affected only decisions of statistical significance and in no way affected actual parameter estimates, which are our major focus.

Statistical Analyses

Because the major focus of this study is on the demonstration of analytical procedures, each procedure is presented here only briefly and is described in greater detail in conjunction with the presentation of the results based on it. For the same reason, we have not pursued further the potentially troublesome issues like the handling of missing data, the use of item-pairs, and the design effects appropriate for the two-stage sampling scheme (but for further discussion, see Bachman, 1970, 1975; Bachman & O'Malley, 1986; Marsh, 1987). In demonstrating these procedures, we used the following outline.

1. We began with a brief evaluation of traditional approaches to the study of individual differences and mean stability. These were based on analyses of scale scores, mean responses averaged across responses to the self-esteem items collected on each occasion. Inferences about the stability of individual differences were based on correlations among the scale scores. Inferences about stability and changes in the mean level of self-concept were based on repeated-measures ANOVAs.
2. Then we explored structural equation models of covariance stability—the stability of individual differences. These analyses were based on covariance matrices for multiple indicators of self-esteem on each occasion instead of single scale scores. As part of these analyses, we examined the invariance of parameter estimates over the multiple occasions. Separate tests were conducted of the invari-

ance over multiple occasions of factor loadings, factor variances, and uniquenesses.

3. Next we described a SEM approach to the examination of latent mean differences. These analyses were based on mean moment matrices that contained information about both covariance stability and mean stability. These models provided a unified approach to the evaluation of the covariance stability and mean stability. Within this approach we illustrated how latent mean differences can be transformed into polynomial trend components and other transformations typically used in repeated-measures ANOVAs (e.g., SPSS, 1988). The choice of such contrasts should be driven by particular substantive issues and may well be made a priori.

4. Finally, we integrate earlier discussion of invariance, demonstrating how interpretations that are warranted depend fundamentally on various levels of invariance of parameter estimates over the multiple occasions. In this context, we develop a hierarchy of invariances in which appropriate levels of interpretation are related to different levels of the invariance of parameter estimates over occasions.

Goodness of fit for all models considered here was evaluated by (a) establishing that a solution converged to a proper solution in which parameter estimates were within the range of permissible values (i.e., there were no negative variance estimates); (b) examining parameter estimates in relation to a priori predictions and common sense; and (c) evaluating the chi square and subjective indices of fit for alternative models. Following Marsh, Balla, and McDonald (1988) and McDonald and Marsh (1990), we used the relative noncentrality index (RNI) and the Tucker–Lewis index (TLI) to evaluate fit, but note that most other indices of fit can be computed from the values presented. TLI and RNI values of .9 or higher are typically used to infer a minimally acceptable fit, but the RNI is monotonic with model complexity, whereas the TLI incorporates a penalty for the inclusion of more parameters. For this reason, we placed more emphasis on the TLI than the RNI for assessing fit.

RESULTS

Traditional Approaches to Covariance and Mean Stability

The traditional approaches to the analysis of both mean and covariance stability begin with a single measure of each construct on different occasions (O1, O2, ... O5; Table 2). Covariance stability typically is inferred on the basis of correlations among measures from different occasions, whereas mean stability is typically inferred on the basis of comparisons of the means from different occasions.

TABLE 2
Self-Esteem Scale Scores for Occasions 1 to 5: Means, Standard Deviations, Coefficient Alpha Estimates of Reliability, and Stability Correlations

				Correlations				
Time	*M*	*SD*	r_{xx}	*O1*	*O2*	*O3*	*O4*	*O5*
1	3.757	.513	.760	–	.687	.596	.535	.387
2	3.835	.482	.770	.525	–	.806	.711	.521
3	3.886	.503	.801	.465	.633	–	.811	.556
4	3.903	.494	.814	.421	.563	.655	–	.614
5	4.220	.477	.812	.304	.414	.448	.499	–

Note. Correlations below the main diagonal are uncorrected; correlations above the main diagonal have been corrected for unreliability. Correlations above the main diagonal have been corrected for unreliability based on coefficient alpha coefficients even though it is subsequently argued that this is not appropriate for multiwave panel studies.

Covariance Stability

Covariance stability is frequently based on an inspection of test–retest correlations among measures collected on different occasions (Table 2). Sometimes, particularly when reliability estimates are available, the test–retest correlations are corrected for unreliability using coefficient alpha estimates of reliability. Inspection of Table 2 reveals that, for the 1-year O2/O3 and O3/O4 intervals, correlations are about .65 before correction for unreliability and about .80 after correction for unreliability. For the 8-year O1/O5 interval, the correlation is .30 before correction for unreliability and .39 after correction. Across all correlations, the mean r is .493 before correction for unreliability and .622 after correction. An inspection of the correlations suggests a simplex pattern (Guttman, 1954; Marsh, 1993) in which self-concept measures that are temporally closer together tend to be more highly correlated, whereas those that are more distant are less highly correlated. Also, there appear to be higher test–retest correlations as respondents become older, although this interpretation is complicated by the unequal length of the intervals—particularly the O4/O5 interval that is 4 years.

Preliminary interpretations of covariance stability based on correlations in Table 2 are plausible, but there are serious problems with their interpretation. In particular, there is inadequate consideration of measurement error and the traditional assumption of independent errors of measurement. For our purposes, we assume that there is true score variance that reflects self-esteem and uniqueness consisting of random error and specific variance particular to the items used to measure self-esteem. In traditional approaches to reliability (e.g., coefficient alpha estimates based on a single set of measures), these two sources of uniqueness are not differentiated. When the same measures are collected on different occasions, however, specific variance

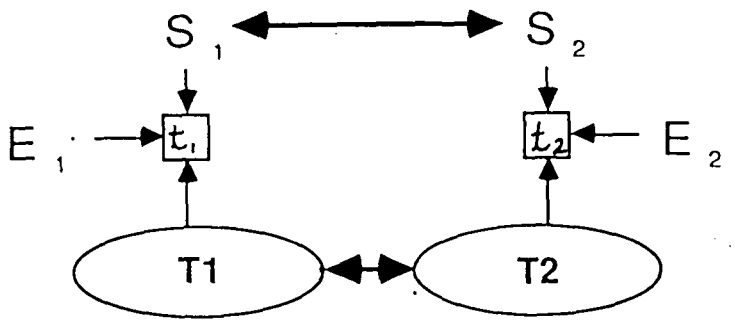

FIGURE 1 Relations between self-esteem on Occasions 1 and 2. The boxes represent observed measures of self-esteem. *T*1 and *T*2 = true self-esteem. *E*1 and *E*2 = random error. *S*1 and *S*2 = specific variance that is idiosyncratic to the instrument.

that is a function of the specific items is likely to have a consistent effect across the different occasions. To illustrate this logic, consider the model in Figure 1 in which observed self-esteem (*t*1 and *t*2; the boxes) is measured at occasions 1 and 2 using the same instrument. Observed responses (*t*1 and *t*2) represent the influences of the true, underlying self-esteem (*T*1 and *T*2), random error (*E*1 and *E*2), and specific variance that is idiosyncratic to the instrument (*S*1 and *S*2). According to the model, the true scores (self-esteem) and specific components are correlated over time, whereas random errors are not.

According to the model in Figure 1, neither the observed nor the disattenuated correlations in Table 2 accurately reflect the correlation between the self-esteem true scores. The observed O1/O2 correlation may have two opposing biases. Because the influence of random measurement error has not been removed, the observed O1/O2 correlation is attenuated and provides a negatively biased estimate of the true correlation. On the other hand, the observed O1/O2 correlation may be positively biased by a failure to control for the positive correlation between *S*1 and *S*2. The model in Figure 1 cannot actually be tested with scale scores, and the information in Table 2 provides no information about the relative strengths of these off-setting biases. The negative bias due to not correcting for error may be larger than the positive bias due to not controlling for correlations between specific variances so that the observed O1/O2 correlation will underestimate the true correlation. In contrast, the O1/O2 disattenuated correlation in Table 2—which has been corrected for coefficient alpha estimates of reliability computed separately at each occasion—is a positively biased estimate of the true correlation. This follows because the effects of error have been removed in the disattenuated correlation, but there is no control for the positively correlated specific components that are confounded with true stability. The focus of structural

equation models to be discussed later is to evaluate more fully these issues in the estimation of correlations between true scores. Although we illustrate these issues with the self-esteem data, concerns about correlated uniquenesses and the potential bias in interpretations of models that exclude them generalize to all multiwave data in which the same indicators are used on different occasions.

Mean Stability

A preliminary inspection of the means (Table 2) suggests that self-concepts show a small, steady increase over the 8-year period. A traditional repeated-measures analysis with polynomial contrasts was conducted that takes into account the unequal intervals (SPSS, 1988). This analysis (Table 3) indicates that there are substantial increases and that nearly all of this increase is linear with time. The major deviation from linearity appears in the O3/O4 interval that represents the first year after graduation from high school. Although self-concept does increase during this interval, the increase is smaller than in other intervals. The deviations from linearity, although very small, are statistically significant due to the very large sample size. The questions to be addressed in subsequent analyses are whether mean differences between latent constructs that have been corrected for error show a similar pattern and whether assumptions underlying the interpretations of the manifest comparisons are plausible.

TABLE 3
Stability of Observed Mean Self-Esteem Over an 8-Year Period:
A Traditional Repeated-Measures Analysis of Polynomial Trend Components

Source	*SS*	*MS*	*F Ratio*[a]
Time (linear)	160.465	160.465	928.90*
Error	226.817	.173	
Time (quadratic)	2.577	2.577	17.93*
Error	188.715	.144	
Time (cubic)	1.517	1.517	15.51*
Error	128.427	.098	
Time (quartic)	0.176	.176	2.16
Error	106.914	.081	

Note. For the polynomial trend analyses, orthonormalized trend components were used that took into account the unequal spacing of the intervals. These were computed by the SPSS-X (1988) package and are presented in Appendix B. For our purposes, each occasion was represented as the number of months since the first data collection: O1 = 1, O2 = 17, O3 = 29, O4 = 42, O5 = 90.

[a]Tests of statistical significance are positively biased due to the stratified random sampling. To compensate for this bias, observed *F* ratios were evaluated in relation to critical values with *df* = 1 and 999, although this had no effect on any decisions summarized in this table.

*$p < .001$.

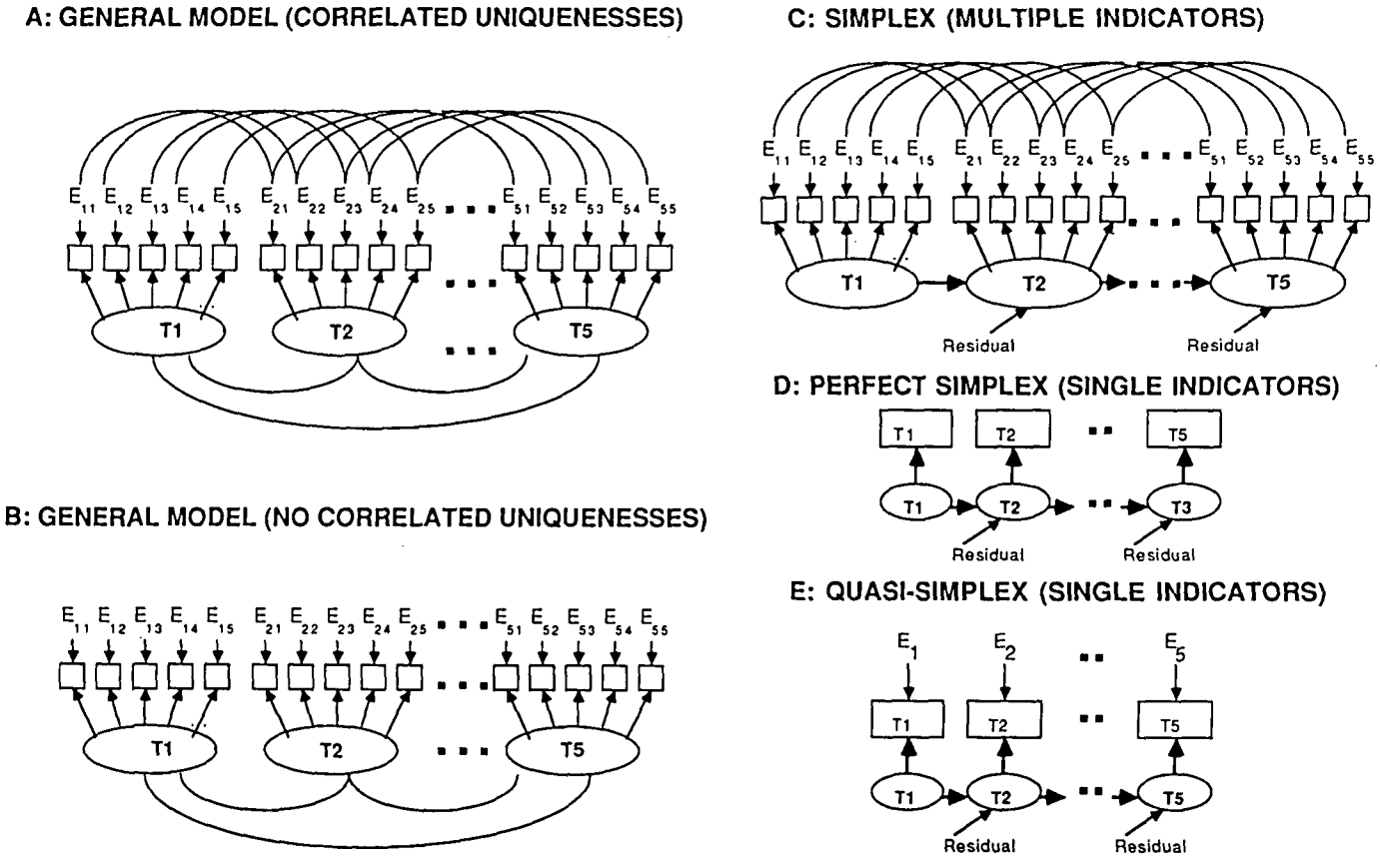

FIGURE 2 Five models of the relations between self-esteem measured with the same items on five occasions. Boxes represent observed measures. Ovals (*T*1 to *T*5) represent latent constructs. *E*s (*E*11 to *E*55) are uniqueness associated with each item on each occasion. Single-headed arrows reflect causal effects, and curved lines reflect correlated variables.

Structural Equation Models of Covariance Stability

General Model for Multiple Indicators.

Structural equation models considered here are intended to evaluate the covariance stability of the same construct measured on multiple occasions. The most general model that we consider is presented in Figure 2A, hereafter referred to simply as the *general model.* In the general model, the same set of five measured variables (the boxes) are used to infer latent constructs (the *Ts in the* ovals) on each occasion. The curved lines connecting the latent constructs represent true score correlations that are of major interest. The *E*s associated with each measured variable represent uniqueness, and the lines connecting the *E*s represent correlated uniquenesses for the same measured variable administered on different occasions. Important goals for such models are to accurately reflect the influences of uniqueness and correlated uniquenesses and to separate these influences from correlations among latent constructs.

Many variations of the general model are possible, such as the elimination of the correlated uniquenesses (Figure 2B), requiring some or all parameter estimates to be invariant over time, and simplex models (Figure 2C) that posit a more parsimonious pattern of relations between the latent constructs. Each of these variations is a special case of (i.e., nested under) the general model in that additional constraints are imposed on the general model. The general model is important because logically it provides more accurate estimates of true test–retest correlations among latent self-esteem constructs than other models that impose additional constraints and because it provides a baseline against which to test these other models. The goodness of fit of this general model sets an upper limit on the fit of alternative (nested) models and thus provides a basis for evaluating the additional constraints imposed in these models. Unless the general model fits the data adequately, it may be unreasonable to pursue additional models.

Parameter estimates based on the general model (Figure 2A) are presented in Table 4. The solution is proper in that none of the parameter estimates fall outside of their permissible region (e.g., negative variance estimates). The latent self-esteem constructs are well defined in that all the factor loadings are large and statistically significant. Whereas the overall chi-square goodness-of-fit test is statistically significant, $\chi^2(215) = 291$, due in part to the large sample size, the TLI (.989) indicates that the goodness of fit is exceptionally good (see Model 1A in Table 5).

Correlated uniquenesses. An important feature of the general model (Figure 2A) is that it includes correlated uniquenesses between the same measured variables presented on different occasions. Inspection of Table 4 indicates that all 50 of the correlated uniquenesses are positive and 44 are

TABLE 4

Parameter Estimates for Multiple-Indicator Model: Factor Loadings, Residual Variances and Covariances of Measured Variables, and Factor Variances and Covariances

	Factor Loading[a]			*Residual Variances and Covariances*[b] *for Measured Variables*				
	Standardized Coefficient	*Unstandardized Coefficient*	*SE*	*O1*	*O2*	*O3*	*O4*	*O5*
Occasion 1								
1	.538	.859	.063	.364				
2	.615	1.000	–	.335				
3	.681	1.158	.074	.319				
4	.615	.895	.061	.269				
5	.634	1.024	.068	.316				
Occasion 2								
1	.560	.931	.063	.097	.323			
2	.627	1.000	–	.040	.265			
3	.765	1.327	.074	.042	.217			
4	.608	.894	.058	.039	.234			
5	.596	.941	.062	.043	.276			
Occasion 3								
1	.605	.896	.052	.087	.111	.299		
2	.686	1.000	–	.035	.022	.245		
3	.748	1.154	.058	.040	.056	.227		
4	.672	.874	.048	.044	.048	.204		
5	.646	.913	.051	.023	.059	.252		
Occasion 4								
1	.644	.965	.057	.057	.082	.094	.243	
2	.632	1.000	–	.005	.037	.018	.279	
3	.789	1.294	.067	.044	.051	.039	.182	
4	.683	.956	.055	.035	.033	.041	.195	
5	.672	.999	.058	.020	.028	.042	.220	
Occasion 5								
1	.622	.921	.056	.038	.047	.062	.057	.235
2	.641	1.000	–	.035	.000	.012	.013	.251
3	.765	1.294	.069	.013	.020	.040	.031	.208
4	.737	1.091	.059	.000	.026	.019	.021	.177
5	.652	.853	.050	.022	.037	.030	.014	.171

(Continued)

TABLE 4
(Continued)

Latent Factor Variances (on Diagonal), Covariances (Above Diagonal), Latent Factor Correlations (Below Main Diagonal)

	O1	*O2*	*O3*	*O4*	*O5*
O1	.204	.118	.115	.097	.068
O2	.632	.172	.143	.118	.086
O3	.546	.742	.217	.152	.101
O4	.500	.661	.760	.184	.104
O5	.362	.496	.517	.578	.175

Note. Results were based on analyses of a 25 × 25 variance–covariance matrix (Appendix A). For this structural equation model, each measured variable was allowed to load on one and only one latent construct, and latent factor variances and covariances were freely estimated. Residual covariances (correlated errors) between each measured variable and the same measured variable from different occasions were freely estimated.

[a]Latent constructs for O1 to O5 are separate constructs, and factor loadings are presented in a single column to conserve space. The factor loading of the second measured variable at each occasion was fixed at 1.0 to set the metric of the latent constructs. To facilitate interpretation, standardized factor loadings and factor correlations based on a parallel analysis of the 25 × 25 correlation matrix are also presented. [b]The residual variance–covariance matrix in the actual analysis was a 25 × 25 symmetric matrix and is presented in this form to conserve space. For each line the last value is the residual variance, and other values are residual covariances (correlated errors) between that measured variable and the same measured variable presented earlier. For example, the residual variance of the fifth indicator at O5 is 0.171, and residual covariances between this indicator and the same indicator at O1 is 0.022.

statistically significant. The first alternative to this general model that we consider is a corresponding model (Figure 2B) in which the 50 correlated uniquenesses are all constrained to be zero. Although the solution resulting from this model is proper, the fit to the data is substantially poorer (see Models 1A and 2A in Table 5). The difference in chi-squares for the two models (969) is very large in relation to the difference in degrees of freedom (50), and the TLIs differ substantially (.880 vs. .989). Other models to be considered subsequently were also fit with and without correlated uniquenesses. Because the inclusion of the correlated uniquenesses consistently had a substantial effect on fit (see Table 5), these additional models are not to be considered further. These results provide strong support for the general model and the need to include correlated uniquenesses.

Comparison of latent and manifest correlations. It is also relevant to compare the 10 correlations among the 5 latent constructs based on these two structural equation models (see Table 6) with the corresponding correlations among scale scores presented earlier. We argued that the failure to take into account correlated uniquenesses would positively bias estimates of the true correlations. Consistent with this claim, correlations based on the gen-

TABLE 5
Structural Equation Models Based on Multiple Indicators: Goodness-of-Fit

Model	χ^2	*df*	*TLI*	*RNI*
Structural equation models of individual differences[a]				
Correlated uniquenesses				
1A 5 correlated first-order factors	291	215	.989	.992
1B first-order simplex	343	221	.982	.987
1C second-order simplex	301	218	.988	.991
1D third-order simplex	291	216	.989	.992
1E fourth-order simplex	291	215	.989	.992
1F 5 uncorrelated first-order factors	1,755	225	.782	.836
No correlated uniquenesses				
2A 5 correlated first-order factors	1,260	265	.880	.894
2B first-order simplex	1,286	271	.880	.891
2C second-order simplex	1,265	268	.881	.893
2D third-order simplex	1,260	266	.880	.894
2E fourth-order simplex	1,260	265	.880	.894
2F 5 uncorrelated first-order factors	3,122	275	.668	.696
Tests of invariance on Model 1A				
1A no invariance (Model 1A)	291	215	.989	.992
3A Model 1A with factor loadings invariant	332	231	.986	.989
3B Model 3A with factor variances invariant	337	235	.986	.989
3C Model 3B with factor covariances invariant	518	244	.964	.971
3D Model 3B with uniqueness invariant	617	255	.954	.961
3E Model 3D with T1 uniqueness free	446	250	.975	.979
3F Model 3D with correlated errors invariant	767	300	.950	.950
3G Model 3F with T1 uniqueness free	562	295	.971	.971
Structural equation models of latent mean structures[b]				
4A 5 correlated first-order factors	760	247	.933	.945
4B Model 4A with polynomial contrast	760	247	.933	.945
4C Model 4A with difference contrast	760	247	.933	.945
4D Model 4A with repeated contrast	760	247	.933	.945
4E Model 4A with all means equal to 0	1,324	251	.863	.885
5A Model 4A with partial invariance	396	244	.939	.953

Note. TLI = Tucker–Lewis index, RNI = relative fit index.

[a]Models were fit to covariance matrices. [b]Models were fit to full-moment matrices. Invariant-means models are constrained such that each measured variable has the same factor loading and the same intercept across all five occasions. For the partially invariant model, this constraint was relaxed. In Model 5A, measured-variable intercepts were freed for three measured variables from O5. Models using the polynomial, difference, and repeated contrasts necessarily have the same goodness-of-fit as the models with five correlated factors but represent a different parameterization of the latent construct means.

TABLE 6
Correlations Among Self-Esteem Latent Constructs at Occasions 1 to 5

Model	*O1/O2*	*O1/O3*	*O1/O4*	*O1/O5*	*O2/O3*	*O2/O4*	*O2/O5*	*O3/O4*	*O3/O5*	*O4/O5*	*Mean r*
Raw scale scores (see Table 2)											
Uncorrected scale scores	.525	.465	.421	.304	.633	.563	.414	.655	.448	.499	.493
Disattenuated scale scores	.687	.596	.535	.387	.806	.711	.521	.811	.556	.614	.622
Multiple-indicator structural equation model[a]											
Correlated factors (1A)[b]	.632	.546	.501	.362	.742	.661	.496	.760	.517	.578	.580
First-order simplex (1B)	.652	.508	.402	.239	.779	.616	.367	.792	.471	.595	.542
Second-order simplex (1C)	.635	.554	.473	.314	.745	.671	.437	.761	.527	.581	.570
Third-order simplex (1D)	.632	.550	.498	.350	.741	.671	.472	.755	.531	.577	.580
Fourth-order simplex (1E)	.632	.546	.501	.362	.742	.661	.496	.760	.517	.578	.580
Multiple-indicator structural equation model[c]											
Correlated factors (2A)	.680	.589	.532	.379	.802	.711	.522	.803	.550	.606	.617
First-order simplex (2B)	.697	.580	.484	.303	.833	.695	.435	.834	.523	.627	.601
Second-order simplex (2C)	.682	.595	.520	.353	.804	.719	.484	.803	.561	.609	.613
Third-order simplex (2D)	.680	.589	.532	.380	.802	.711	.522	.803	.550	.606	.617
Fourth-order simplex (2E)	.680	.589	.532	.379	.802	.711	.522	.803	.550	.606	.617
Invariance model[d]											
Factor loadings invariance (3A)	.633	.546	.501	.363	.743	.661	.495	.760	.519	.578	.580
3A with F variances invariance (3B)	.633	.545	.500	.364	.741	.657	.495	.757	.518	.578	.580
3B with F covariance invariance (3C)	.589	.589	.589	.589	.589	.589	.589	.589	.589	.589	.589
3B with uniqueness invariance (3D)	.613	.535	.494	.360	.742	.666	.506	.770	.530	.596	.580
3C with corr unique invariance (3E)	.610	.532	.484	.336	.763	.679	.495	.785	.532	.599	.581

[a]Correlated errors; see Table 5. [b]Model 1A provides the most accurate estimate of correlations among latent constructs at different occasions in that Models 1A to 1E, 2B to 2D, and 3A to 3E are all special cases of Model 2A in which additional constraints are added. To the extent that any of these additional constraints alter correlations based on Model 2A, the alternative model is inaccurate. Standard errors for each of the different correlations are about .03. [c]No correlated errors; see Table 5. [d]See Table 5.

eral model with correlated uniquenesses (mean $r = .580$; see model 1A in Table 6) are systematically lower than corresponding correlations for the general model without correlated uniquenesses (mean $r = .617$; see model 2A in Table 6).

We also argued that observed correlations among scale scores may underestimate true correlations, whereas observed correlations corrected for coefficient alpha estimates of reliability would overestimate the true correlations. In support of these claims, all 10 correlations among latent constructs for the general model (mean $r = .580$) fall between the corresponding correlations based on scale scores (mean $r = .493$) and the corrected correlations among scale scores (mean $r = .622$). We argued further that correlations based on the multiple-indicator model without correlated uniquenesses are conceptually similar to scale score correlations that are corrected for coefficient alpha estimates of reliability. Consistent with this claim, the 10 correlations based on the general model with no correlated uniquenesses (mean $r = .617$) are consistently similar to the corrected correlations among scale scores (mean $r = .622$). This empirical support for a priori predictions provides further support for the superiority of the general model with correlated uniquenesses and the logical basis for understanding the systematic biases in correlations based on other approaches.

Simplex model. Guttman's (1954; see also, Jöreskog, 1970, 1979) classic study of the simplex structure for relations among ordered tests provided an alternative representation that is one of the most popular approaches to evaluating covariance stability (Marsh, 1993). The most salient property of the simplex for longitudinal data is that the sizes of correlations between measures collected at adjacent occasions are largest and that the sizes of correlations steadily decrease as a function of the number of occasions separating two measures. In a first-order simplex model (see Figures 2C), self-esteem at each occasion is posited to directly affect self-esteem on the next occasion, but not to directly affect self-esteem on subsequent occasions. Thus, the correlation between self-esteem at any two nonadjacent occasions (e.g., O1 and O3) is zero when self-esteem from an intervening occasion (e.g., O2) is partialled out. In a second-order simplex, self-esteem at each time is posited to directly affect self-esteem on the next two occasions. Given a sufficient number of occasions, it is possible to fit third-order, fourth-order, or even higher orders of simplex models, although most researchers have focused on the first-order simplex.

Marsh (1993; also see O'Malley & Bachman, 1983) argued for the superiority of the multiple-indicator simplex model considered here (Figure 2C) over the single-indicator simplex model based on scale scores. The multiple-indicator simplex model is also nested under the general model (Figure 2A); it is just a special case of the more general model with the additional

assumption that correlations among the latent constructs form a simplex. If the simplex model accurately reflects the data, then the corresponding general model—because the simplex model is nested under it—must also accurately reflect the data, but the converse is not true. Whereas the first-order simplex is nested under the more general model, a simplex model with a sufficiently high order (i.e., paths connecting nonadjacent constructs differing by 2, 3, 4, or more occasions) is mathematically equivalent to our general model. The highest order simplex model that can be fit is one less than the number of occasions for the multiple-indicator models. Here, for example, the 4th order simplex is equivalent to our general model.

The first-order simplex model (Model 1B in Table 5) provides a significantly poorer chi-square than the general model (Model 1A in Table 5); the difference in chi-squares is 52 in relation to a difference of 6 in degrees of freedom. Nevertheless, the TLI for the simplex model (.982) is very good and only marginally lower than for the general model (.989). Much of this overall ability to fit the data, however, is based on the good fit of the basic measurement model (e.g., the first-order factor loadings, measurement errors, and correlated errors) that is common to both the general and simplex models. For this reason, it is also important to evaluate the correlations among the self-esteem constructs based on this simplex model. (Although the correlations between latent constructs are not actually estimated in the simplex model, they are presented as part of the output of most SEM packages.)

Correlations based on the general model (mean $r = .580$; see Model 1A in Table 6) are larger than correlations based on the simplex model (mean $r = .542$; Model 1B in Table 5), but there is a systematic pattern of differences. Correlations based on the simplex model are consistently too high for temporally adjacent constructs and consistently too small for temporally nonadjacent constructs. The largest difference is between self-esteem at O1 and O5, that is .362 in the general model and .239 in the simplex model. Not surprisingly, correlations among the latent constructs based on the simplex model are biased in the direction of a simplex pattern. This apparent bias in the first-order simplex model is substantially reduced in the second-order simplex model (1C), almost completely eliminated in the third-order simplex model (1D), and completely eliminated in the fourth-order simplex model (1E) that is merely a reparameterization of the general model (1A). Although not a major focus of this investigation, problems with the simplex model such as these are evaluated in greater detail by Marsh (1993) and seem to seriously undermine the usefulness of the simplex model.

Factorial Invariance Based on the Covariance Matrix

A very difficult problem in longitudinal studies of the same construct over time is that psychometric properties of the measurement instrument

may change. Whereas it may be reasonable to assume the invariance of these properties over short intervals, this assumption becomes more problematic as the time intervals become longer, and particularly when psychological constructs are inferred during a developmental period in which change is expected. A preliminary inspection of Table 2, for example, reveals that, whereas coefficient alpha estimates of reliability are reasonably stable over occasions, the estimate is somewhat lower at O1. Also, the variability of responses at O1 is somewhat larger than at subsequent occasions. These two observations imply that error variance is larger at O1. Consistent with these observations, the uniquenesses in the general model (see Table 4) are systematically larger at O1 than at other occasions. In contrast, factor loadings appear to be reasonably consistent over time.

The problem of testing the invariance of a structure for a single group over multiple occasions is analogous in many ways to the more frequently studied problem of testing the invariance of parameters across different groups (e.g., Byrne et al., 1989; Jöreskog & Sörbom, 1988; Marsh, 1994; Marsh & Hocevar, 1985). In the multiple group approach it is possible to set any one parameter, any set of parameters, or all parameters to be equal in multiple groups. To the extent that a model with invariance constraints is able to fit the data as well as a corresponding model without constraints, there is support for the invariance constraints. These studies typically test for the invariance of different sets of parameters—particularly factor loadings, but also factor variances and covariances and measurement errors. Marsh and Hocevar argued that the minimum condition for factorial invariance is the invariance of factor loadings. Byrne et al., however, argued for tests of partial invariance in which it is only necessary for at least one estimated factor loading for each latent construct to be invariant across groups. In analogous tests of the invariance of the same measure over multiple occasions, it makes sense to test for the invariance of factor loadings over multiple occasions, but also, perhaps, of the true score variances of latent constructs and of the measurement errors for the same measured variable. As in the multiple group situation, the minimum condition for factorial invariance would seem to be the invariance of factor loadings over time, although it may be useful to pursue Byrne et al.'s notion of partial factorial invariance when tests of factorial invariance fail.

The goodness of fit for alternative tests of invariance is summarized in Table 5. Model 1A (Figure 2A), the general model with correlated uniquenesses, must necessarily be able to fit the data as well as any model that imposes invariance constraints, and so it provides an important basis of comparison. In the first test of invariance (Model 3A), factor loadings for the same measured variable are held constant across the five occasions. Although the change in $\chi^2 = 41$ is statistically significant in relation to the change in $df = 16$, due in part to the large sample size, the goodness-of-fit indices are nearly the same. We interpret this as support for the

invariance of the factor loadings over time. In the second test of factorial invariance (Model 3B), the factor variances of the five latent constructs—*self-esteem* at the five occasions—are tested. The change in $\chi^2 = 5$ in relation to the change in $df = 4$ is not statistically significant, and so there is also support for this invariance constraint. When uniquenesses are held invariant over time (Model 3D), however, the change in $\chi^2 = 280$ due to this additional constraint is substantial in relation to the change in $df = 20$, and the goodness-of-fit indices are poorer. We interpret these results to mean that measurement errors are not invariant over time. Inspection of the modification indices indicated that measured variables from O1 contributed substantially to the lack of invariance of measurement errors over time, and this is consistent with earlier observations that error variance was larger at O1. In Model 3E, measurement errors were constrained to be invariant over O2–O5, but not O1. This substantially improved the fit, although the resulting model still did not fit as well as Models 2A, 3A, or 3B. Hence, these tests provide reasonable support for the invariance of factor loadings and factor variances over occasions, but not measurement error.

Structural Equation Models of Latent Mean Differences

In introducing the problem of testing latent mean structures over multiple occasions, it is again useful to consider the more frequently studied problem of testing the invariance of mean structures and latent means across multiple groups (e.g., Byrne et al., 1989; Jöreskog & Sörbom, 1988; Marsh & Grayson, 1990). In this section we fit several useful latent mean models to the self-esteem data. The distinctions among these models are based in part on issues of invariance. The substantive implications of these tests of invariance are developed further in the next section in which we introduce a hierarchy of invariances.

Analyses of latent mean differences differs from the earlier comparisons of covariance structures in several important ways. First, the starting point for tests of the invariance of the covariance structures is the covariance matrix that is independent of the means of measured variables so long as the variables have a multivariate normal distribution. In contrast, the starting point for tests of mean structures is the matrix of moments about zero that incorporates the mean values. Second, there are two additional vectors of parameters to be estimated: a vector of regression intercepts for the measured variables and a vector of latent means that is of principal interest. In the typical multiple-group problem (e.g., Jöreskog & Sörbom, 1988; Sörbom, 1974), factor loadings and intercepts for measured variables are assumed to be invariant over groups, the latent means for one group are specified to be zero, and the latent means for other groups are freely estimated so that all latent means are scaled in relation to latent means from the

first group. Byrne et al. (1989), however, argued that it still may be reasonable to compare latent means so long as there is at least partial invariance. Subsequent analyses can be used to specify that one, some, or all of the latent means are invariant across groups and to test specific patterns of latent mean differences.

The problem of testing latent mean structures based on a single group over multiple occasions is analogous to testing mean structures across multiple groups. In this investigation, for example, factor loadings and intercepts for measured variables are assumed to be invariant over occasions, the latent mean for O1 is fixed to be zero, and latent means for other occasions are freely estimated so that all latent means are scaled in relation to the latent mean at O1. In subsequent analyses, latent mean differences can be transformed so that estimated latent parameters represent polynomial trend components (e.g., linear, quadratic, cubic components) of growth in means over time as in the typical repeated-measures analysis (see Appendices B and C) or to represent other ANOVA contrasts.

The data interpretations that are warranted are of central importance in the evaluation of latent mean structures, and depend fundamentally on tests of the invariance of parameters over time. In this context it is useful to adopt terminology from item response theory. We do not develop the formal equivalence between item response theory and SEM more fully because we are merely using terminology from item response theory by way of analogy and because this equivalence is demonstrated elsewhere (Christofferson, 1975; McDonald, 1982, 1985). In this approach, each measured variable *(t)* is related to the latent construct *(T)* by the equation $t = a + bT$. The *b* parameter is the slope (or discrimination) parameter that reflects how changes in the observed variable are related to changes in the underlying construct. In structural equation models, the factor loadings reflect the *b* parameter, and support for the factorial invariance of factor loadings over time implies that the *b* parameters are invariant over time. The *a* parameter is the intercept (or difficulty) parameter that reflects the ease or difficulty in getting high manifest scores for a particular measured variable. The comparison of mean stability over time is most justifiable when both the *a* and *b* parameters are invariant over time. When both the relation between a measured variable and the latent construct (the *b* parameter) and the difficulty of getting a high score (the *a* parameter) are invariant over time, then changes in the measured variable apparently imply changes in the underlying construct. Typically, structural equation models are based on covariance matrices that contain no information about the means of the measured variables. In this instance it is possible to test the invariance of factor loadings (the *b* parameters) over time as we did earlier. For structural equation models based on moment matrices that also contain information about the means, however, it is possible to test the invariance of both the factor loadings (the *b* parameters) and the measured variable intercepts (the *a* parameters).

General Latent Means Model: Investigating Latent Trends

Although alternative approaches to the analysis of moment matrices are incorporated in different statistical packages, we chose what appears to be the most widely applied approach—the LISREL approach based on augmented moment matrix—described in the LISREL manuals (Jöreskog & Sörbom, 1985, 1988) and elsewhere (e.g., Bollen, 1989; Byrne et al., 1989; Marsh & Grayson, 1990; Millsap & Everson, 1991). In this approach: (a) the augmented moment matrix (computed by LISREL from the means, standard deviations, and correlation matrix in Appendix A) is a raw product–moment matrix augmented by one row and column (the 26th) to accommodate a constant variable consisting of the manifest means for the 25 measured variables and a 1 in the diagonal; (b) the original 25 measured variables (5 Measured Variables × 5 Occasions) are specified to be y variables, and the associated 5 latent constructs are specified to be ηs; (c) the constant variable is given a fixed-X specification; (d) the factor loading matrix Λ_y is augmented by one column (the 6th) to accommodate the constant variable, which is used to represent the measured variable intercepts; (e) the factor covariance matrix Ψ is augmented by one row and column (the 6th) of fixed zeros to accommodate the constant variable; (f) the mean differences in latent constructs are estimated in Γ by fixing the parameter estimate for O1 to be zero and freely estimating parameters for the other four occasions. In subsequent models, various transformations of the mean differences (e.g., polynomial trends) are introduced to facilitate interpretations of the means (see Appendix B for the LISREL setup and Appendix C for technical information on the matrix formulation used to generate this LISREL setup; also see Bock, 1985, for a discussion of the estimation of chosen trends in longitudinal data, or Rao, 1973, for further information). To maintain comparability with earlier analyses, we used the same null model as in analyses based on the covariance matrix. It may, however, be appropriate to use a more restrictive null model for analyses of mean moment matrices in which invariance constraints are placed on the intercepts of each measured variable consistent with those in subsequent models. This would result in slightly better fit indices for all latent mean models presented here so that our decision is conservative in relation to obtaining apparently acceptable fit indices.

Tests of the general latent mean model. For the set of invariant means models considered here (Models 4A to 4E in Table 5), factor loadings and the intercepts of the measured variables were fixed to be invariant over multiple occasions, and the mean of the latent self-esteem construct was fixed to be zero for O1. Model 4A, which posits five correlated factors, resulted in a substantial chi-square (Table 5), but the TLI indicates that the

fit is good and only modestly poorer than the TLI for the corresponding Model 3 based on the covariance matrix. Whereas the fit indices apparently support acceptance of the model, the large chi-square dictates that this needs to be interpreted cautiously. For this reason and for purposes of demonstration, we chose to pursue two different courses. First, we interpret results from the invariant means model and fit additional models with the same invariance constraints based on the assumption that the fit indices justify support of these invariance constraints. Second, we pursue the notion of partial invariance by freeing selected parameters that are invariant in the Model 4A and then evaluate parameter estimates based on this a posteriori, partially invariant structure.

Evaluation of latent means. The critical parameter estimates in the means model are the latent mean differences shown in Table 7. For invariant mean models, the latent mean for O1 is fixed to be zero, and means for subsequent occasions are freely estimated. In relation to the standard errors of the estimates, estimates for O2–O5 each differ significantly from self-esteem at O1. The increase in means is monotonic and primarily linear (taking into account the large interval between occasions 4 and 5). This initial interpretation is very similar to that based on the manifest mean scores. (To emphasize the similarity of procedures used to compare latent means with the corresponding procedures based on manifest—raw score—means, both sets of means are presented in Table 7 for a number of models.) We now illustrate how contrasts like those typically used in ANOVA can be used to test specific comparisons among the latent means.

Latent mean trends. In Model 4A, the latent mean for O1 was fixed to zero, and each subsequent latent mean estimate represented the difference between that latent mean and the O1 estimate. This representation of means is the same as the simple contrast (SPSS, 1988) in ANOVA in which the mean of one variable—O1 self-esteem in this case—is compared to all other means. A comparison of the mean difference estimates and the standard errors based on the SEM analysis of latent mean structures with the corresponding values based on a repeated-measures ANOVA of manifest means using the simple contrast (Table 7) indicates that the corresponding estimates are very similar. (This is largely because the five indicators used here function essentially as parallel forms, but this need not always be the case.)

A useful feature of the repeated-measures approach to longitudinal data is the ability to formulate contrasts of particular interest. The polynomial trend contrast (see Table 3), for example, is particularly useful for longitudinal data. Using the same contrast coefficients as in a traditional ANOVA, it is possible to transform latent mean estimates in the analyses of mean structures to represent contrasts of particular interest (see Appen-

TABLE 7
Estimates of Mean Self-Esteem at Occasions 1 to 5 and Transformations of the Means

	Parameter Estimates and Standard Errors									
Model	*O1*[a]	*SE*[b]	*O2*	*SE*	*O3*	*SE*	*O4*	*SE*	*O5*	*SE*
Raw scale scores	3.757	.016	3.835	.015	3.886	.016	3.903	.016	4.220	.015
O1 versus Each Subsequent O	*Constant*	*SE*	*O2–O1*	*SE*	*O3–O1*	*SE*	*O4–O1*	*SE*	*O5–O1*	*SE*
Repeated-measures simple contrast	3.920	.011	0.079	.013	0.130	.015	0.146	.015	0.463	.016
Invariant model (4A)	0.000		0.079	.015	0.132	.016	0.152	.016	0.452	.019
Partial invariant model (5A)	0.000		0.082	.015	0.134	.016	0.157	.017	0.378	.021
Polynomial Trend Components[c]	*Constant*	*SE*	*Linear*	*SE*	*Quadratic*	*SE*	*Cubic*	*SE*	*Quartic*	*SE*
Repeated-measures polynomial contrast	8.766	.023	0.349	.011	0.044	.010	0.034	.009	0.012	.008
Invariant polynomial model (4B)	0.000	.000	0.341	.014	0.035	.011	0.030	.009	0.011	.009
Difference Contrasts[d]	*Constant*	*SE*	*O1 vs. O2*	*SE*	*O1 + O2 vs. O3*	*SE*	*O1 + O2 + O3 vs. O4*	*SE*	*O3 + O4 vs. O5*	*SE*
Repeated-measures difference contrast	3.920	.011	0.079	.013	0.090	.011	0.077	.011	0.375	.012
Invariant difference model (4C)	0.000		0.079	.015	0.092	.012	0.082	.012	0.361	.015
Repeated Contrasts[e]	*Constant*	*SE*	*O2–O1*	*SE*	*O3–O2*	*SE*	*O4–O3*	*SE*	*O5–O4*	*SE*
Repeated-measures repeated contrast	3.920	.011	0.079	.013	0.051	.012	0.017	.011	0.317	.013
Invariant difference model (4D)	0.000		0.079	.015	0.052	.013	0.021	.012	0.300	.016

[a]Constant term is the O1 mean for the raw scale scores, the mean of O1 to O5 for all the repeated measures, and was fixed to zero in all the models. [b]Because the constant term was fixed in the model analyses, no standard error was estimated. All other standard errors were based on a nominal sample size of 1,000. [c]For all polynomial trend analyses, orthonormalized trend components were used based on the number of months since the first data collection: O1 = 1, O2 = 17, O3 = 29, O4 = 42, O5 = 90. [d]In the difference contrast, each score is compared to the mean of all the scores that proceed it. [e]In the repeated contrast, each score is compared with the score that comes next.

dices B and C). For example, using the same orthonormal contrasts as in the repeated-measures analysis (see Table 3), the five latent means in Model 4A were transformed into values representing a constant term (arbitrarily set to be zero for purposes of identification) and the linear, quadratic, cubic, and quartic trend components of time. It is important to emphasize that this model is merely a transformation of the initial means model so that the goodness of fit is not affected (see Table 5). The linear trend component for the latent means is very large and statistically significant, the quadratic and cubic components are substantially smaller but still statistically significant, and the quartic component is not statistically significant. This pattern of results is again similar to that observed in the repeated-measures analysis of manifest scale scores (Table 7) for data considered here. (This similarity depends in part on the multiple indicators acting as parallel forms of the underlying construct.)

Results for other contrasts applied to latent and manifest means, in addition to the simple and polynomial contrasts already discussed, are also presented in Table 7. The difference contrast (SPSS, 1988) is the difference between each mean and all prior means, for example, O2 versus O1; O3 versus (O1 + O2)/2, and so forth. This contrast shows that self-esteem at each occasion is significantly higher than the mean of self-esteem measures from earlier occasions. The repeated contrast (SPSS, 1988) is the contrast between each mean and the subsequent mean (e.g., O1 vs. O2; O2 vs. O3; etc.). This contrast shows that self-esteem increases significantly during each interval except the O3/O4 interval that represents the first year after graduation from high school. For each of these contrasts, the latent means in Model 4A are merely transformed so that the number of estimated parameters and goodness of fit are not altered. Hence, the similarity of the estimates based on latent means to the manifest (raw score) means observed earlier for these data is necessarily observed in the results for each of these contrasts as well.

Partially Invariant Means Model

Despite the apparently good fit indices for the invariant means model (4A in Table 5), the large chi-square dictates caution in the subsequent interpretations. For this reason and for purposes of demonstration, we evaluated a partially invariant means model. There are, however, potential problems with this approach, and some readers may find even the logic of partial invariance to be questionable, or at least debatable. Models based on this approach are a posteriori, and this new approach has not been widely applied. More important, the partially invariant approach is based on the apparently questionable premise that the underlying latent scale is accurately reflected by those indicators that are invariant over time. For these reasons, we agree with the Byrne et al. (1989) suggestion that results from this approach should also be interpreted cautiously. From this perspective, it is interesting to compare

the results based on the invariant and partially invariant approaches. Using an ex post facto approach based on an inspection of LISREL's modification indices (Jöreskog & Sörbom, 1988), we evaluated invariance constraints for the factor loadings and measured variable intercepts over the 5 occasions. Inspection of LISREL's modification indices indicated that relaxing invariance constraints for O5 intercepts would have the largest effect on the chi-square. A step-wise process in which we freed the parameter associated with the highest modification index at each successive step resulted in freeing three of five measured variable intercepts for O5 (Model 5A in Table 5). In this Model 5A, the latent means for O1 to O4 are similar to those in the totally invariant means model (4A), whereas the latent mean for O5 is moderately smaller (Table 7). Thus, the interpretation of the results of Model 5A is still similar to that offered for the invariant means model (4A). Although not pursued here, it would also be possible to transform these mean differences using polynomial, difference, and repeated transformations like those applied to the corresponding totally invariant means Model 4A (see Appendices B and C, Table 7).

Substantive Interpretations Justified by Different Levels of Invariance: A Hierarchy of Invariance

So far, we have discussed latent models of covariance stability and latent mean stability. The issue of the invariance of parameters over time is fundamentally related to the substantive interpretations of such models. Hence, it is important to evaluate limitations imposed on substantive interpretations when these tests of invariance over time fail. Here we develop a hierarchy of invariances and the corresponding interpretations that are justified for different patterns of invariance. If we are interested in evaluating the stability of means (at the latent level), we begin with the full moment matrix of manifest indicators. This will be a function of the psychometric slope (i.e., factor loadings, the *b* parameter in item response theory) and difficulty parameters of each indicator (manifest indicator intercepts, the *a* parameter in item response theory) and the latent moment matrix of the latent constructs. If we are interested in evaluating the stability of individual differences (at the latent level), we begin with the covariance matrix of multiple manifest indicators. This will be a function of the psychometric slope parameters and the latent covariance matrix of the latent constructs.

An implicit assumption underlying the structural equation models considered here is what we refer to as *construct invariance.* We must assume that we are in fact inferring self-esteem at each occasion rather than self-esteem at O1, locus of control at O2, mathematical achievement at O3, and so forth. None of the interpretations of the data discussed here are justified without this axiomatic assumption of construct invariance. Construct invariance does not require any parameters to be invariant over

time, but the assumption of construct invariance seems to be more plausible when this invariance does exist. Even with perfect invariance of factor loadings and manifest intercepts in structural equation models based on moment matrices, it is logically possible that the nature of the construct we are testing has shifted over time, making scores (manifest and latent) incomparable at a fundamental level. It is difficult to see how this could be tested, and it seems that construct invariance must ultimately be assumed in the manner of an underlying axiom. However, better support for the psychometric invariance of parameters in a given data set apparently justifies greater faith in this underlying axiom.

In analyzing the full manifest moment matrix, the ideal situation is the *complete psychometric invariance* of slope and the intercepts of the manifest indicators. In this ideal case, both the latent covariance matrix and the latent mean differences are apparently interpretable. Results presented here provide reasonable support for this level of invariance (Table 5), thus supporting interpretations of latent mean differences.

A somewhat less satisfying situation is *partial psychometric invariance.* This partial invariance, however, can occur in different ways. As Byrne et al. (1989) discussed, some of the indicators may maintain the same a and b values over multiple occasions (i.e., the same ratios or profiles over the occasions; they can then be rescaled to be invariant over occasions by altering the latent moments accordingly). In this case, an apparently plausible argument can be presented that the other indicators are acting in a psychometrically erratic manner and the psychometrically invariant indicators can be trusted to impose a latent scaling so that the interpretation of latent mean differences is still meaningful. Such an argument apparently has more appeal when, say, four of five (rather than just one of five) indicators maintains full psychometric invariance.

A different form of partial psychometric invariance may arise when the slope parameters (i.e., factor loadings) are invariant over occasions, but the difficulty parameters (manifest intercepts) are not. This lack of invariance suggests that the difficulty of the items may have changed so that the comparison of latent means is dubious. This is the situation referred to as factorial invariance (of factor loadings in the covariance metric). In this case, the latent covariance matrix is available for substantive interpretation, even though we have to surrender interpretation of the means because of the psychometric instability of the difficulty parameters. Note that if we had only single indicators at each occasion, such data typically would be analyzed with a repeated-measures ANOVA in which there is hopeless and unknown confounding in the means of genuine construct changes over time with nuisance difficulty parameter shifts.

In this situation, when the interpretation of means is not viable, the slope parameters may still be invariant and the covariance data meaningfully interpreted. When we have factorial invariance (of factor loadings in the covariance metric), we have a latent covariance matrix available for interpre-

tation. Useful hypotheses can be investigated at this level. For example, we may wish to investigate whether the development of self-esteem during adolescence exhibits the typical fanning effect in which self-esteem becomes more dispersed as individuals become older. (Walberg & Tsaai, 1983, presented data showing that smaller differences at younger ages are associated with larger differences at older ages for a wide variety of variables—a pattern that he called the Mathew Effect using the biblical analogy of the rich getting richer and the poor getting poorer.) The nonsignificant difference (Table 5) between Models 3A (factor loadings invariant) and 3B (factor loadings and factor variances invariant) provides a test of this hypothesis. The results indicate that variance in latent self-esteem is quite stable during the adolescent period. (It is also worth noting that tests of such hypotheses based on manifest variables typically confound true latent construct variance and uniqueness.)

A worse scenario would be a complete lack of invariance in which both slope (factor loadings) and manifest intercepts varied over time. In this case, we still have the latent correlation matrix available for interpretation, conceptually derived from the latent z scores for self-esteem at each occasion, so long as the model is able to fit the data. This may still be of substantive value. For example, suppose our data had demonstrated a clear failure of factorial invariance and we had obtained the latent correlation matrix (by fixing latent variances at unity and allowing completely free—noninvariant—factor loadings) reported in Table 6 for Model 2A. Even here, we are justified in applying the same polynomial contrasts as used in the analysis of latent means (see Appendices B and C). Applying this transformation (or any other full rank normalized transformation) to the latent correlation matrix results in a covariance matrix of trend components representing the constant (mean), linear, quadratic, cubic, and quartic trends applied to z scores for each occasion (Table 8). It must be emphasized that these interpretations of

TABLE 8
Covariance Matrix of Polynomial Trend Components of Individual Differences in Latent z Scores

Component	*Constant*	*Linear*	*Quadratic*	*Cubic*	*Quartic*
Constant	3.32				
Linear	−0.10	0.67			
Quadratic	−0.23	−0.32	0.50		
Cubic	0.07	0.00	−0.03	0.30	
Quartic	0.02	−0.01	−0.02	−0.01	0.21

Note. The covariance matrix presented here is a simple transformation of the correlations among latent constructs presented in Table 6 based on the polynomial transformation described in Appendix B. The variance terms in the diagonal reflect the extent to which subjects vary on each component, whereas the covariance terms indicate the extent to which individual differences in one component are related to individual differences in other components.

these values are limited to (z score) individual differences among subjects and have nothing to do with changes in group means (which are necessarily zero when z scores at each occasion are considered). Nevertheless, substantively important hypotheses can still be evaluated in relation to both trend variance and covariances.

In Table 8, the variance of the constant (mean) term is much larger than any of the remaining polynomial trend components. It seems that the rank ordering of subjects remains stable over time—subjects who have high z scores at any one occasion tend to have high z scores on all occasions. We interpret this result as strong evidence against the storm and stress hypothesis in which self-esteem would be expected to become increasingly more stable over the adolescent-to-young adult period considered here. An alternative pattern might have been one in which the rank order of some subjects systematically increased (z scores became more positive) whereas other subjects systematically decreased (z scores became more negative). This pattern would have resulted in a large variance in the linear trend component and would have provided support for the storm and stress hypothesis. This pattern of results would also have suggested the usefulness of exploring other variables (e.g., sex, socioeconomic status, school grades) that may differentiate among those students with positive and negative trend components. The relatively small variance component associated with the linear trend component, however, indicates that subjects did not differ substantially on the linear trend component, and so this subsequent analysis would probably not be particularly useful.

It is also useful to evaluate the covariance terms among the different trend components. If, for example, the constant (mean) term were positively correlated with the linear term, it would imply that rank order of subjects who had high z scores at O1 tended to have increasing z scores over time, whereas subjects who had low z scores at O1 tended to have decreasing z scores over time. Inspection of Table 8, however, indicates that the constant mean term is relatively uncorrelated with any of the other trend components. In fact, with the possible exception of the correlation between the linear and quadratic trends, all the covariance terms are small. Even this one apparently nonnegligible covariance is not important, because neither the linear nor the quadratic trend components account for much variance.

In this investigation, because we had factorial invariance (invariance of factor loadings in the covariance metric) and invariant latent construct variances, the latent covariance matrix is approximately the same as the latent correlation matrix (up to a single multiplicative scaling constant). Thus, the same conclusions could be drawn at the level of latent scores in a stable interval metric as well as at the level of z scores. So, in this context, we can say that subjects entering the study with relative self-esteem that was low (or high) tended to remain so throughout.

SUMMARY AND DISCUSSION

The substantive importance of this investigation was to provide a rigorous evaluation of the stability of self-esteem during adolescence. The results demonstrated that levels of self-esteem (latent means) and the stability of individual differences (test–retest) correlations increased during this period. Subjects who began with relatively high (or low) self-esteem at O1 tended to have high (or low) self-esteem across all five occasions. There was no indication of dramatic upheavals in self-esteem that would support a period of stress and storm. In addition to these substantively important findings, this investigation provided a methodological demonstration that should have broad applicability to the study of individual differences in multiwave, longitudinal data.

Individual Difference Stability

The results of this investigation have clear implications for the methodology of studies of individual difference stability. Marsh (1993) and O'Malley and Bachman (1983) argued for the use of a multiple-indicator simplex model instead of simplex models based on single indicators whenever multiple indicators are available. Here, we further demonstrated—logically and empirically—that the general model (Figure 2A) is even better for inferring correlations among latent traits. This conclusion is important because each of the other approaches considered here can be considered as special cases of this general approach that require additional assumptions. If these assumptions are reasonable, then the alternative approach should result in estimates similar to the general model and may be argued to offer a more parsimonious description of the results. If these results are not reasonable, however, the alternative approaches will provide systematically biased estimates. This investigation indicated that each of the alternative approaches provided biased estimates for the data considered here, and at least the direction of these biases was predicted a priori. Whereas the size of the biases may vary substantially from study to study, the direction and pattern of biases observed here apparently will generalize broadly.

Correlations based on raw scale scores were negatively biased, whereas correcting these correlations for coefficient alpha estimates of unreliability resulted in positively biased estimates. Multiple-indicator models that did not take into account the correlated errors also resulted in positively biased estimates of these correlations, and the size of the bias was about the same as for scale score correlations corrected for coefficient alpha estimates of unreliability. Although the multiple-indicator simplex model provided positively biased estimates when correlated errors were ignored and slightly negatively biased estimates when correlated errors were included, there was a systematic pattern in which relations between temporally adjacent con-

structs were more positively biased whereas relations between temporally more distant constructs were more negatively biased.

Latent Mean Approaches

Our results also have clear implications for the application of latent mean models with longitudinal data. This investigation demonstrates some potentially important advantages to the SEM analytic approach. The ability to make inferences on means of latent constructs instead of manifest variables is appealing, particularly when it is possible to conduct further tests using the array of contrasts that are applied in the typical ANOVA approaches. There are, however, problems that require further research before this approach can be routinely applied.

In the application presented here, the metric of the mean parameters was determined by fixing the factor loadings for one of the measured variables for each construct to be 1.0 and fixing the first latent mean to be zero. Had we chosen to fix the first latent mean to another constant—for example, the mean of the corresponding manifest scale score—all of the parameter estimates in Model 4A would have differed by an additive constant. We chose the second indicator of the latent means to serve as a reference indicator because its factor loading was the most central—there were two factor loadings that were larger and two that were smaller when factor loadings were held invariant over occasions—and in part because the modification indices indicated that holding this indicator invariant over occasion had the least serious implications for goodness of fit. This choice also tended to maximize the similarity of values based on latent and manifest variables. If we had chosen another variable to serve as the reference indicator, all the parameters in Model 4A would have differed, although the ratio of estimates for the two different scalings and corresponding t values (parameter estimates divided by standard errors) would have been the same. This degree of arbitrariness is of no substantive importance in the fully invariant means model but may be troublesome for the partially invariant means model.

The implications of misfit models also need further evaluation. In the application presented here, there was good support for the invariance of factor loadings based on the covariance matrix (e.g., Model 3A in Table 5), but somewhat weaker support for the invariance of factor loadings and measured variable intercepts based on the mean moment matrix. The relative fit indices suggested that the fit for the model based on the moment matrix is acceptable, but there is need for further study to establish guidelines about what constitutes an acceptable fit in this situation. Because we were concerned about the fit of the model in which factor loadings and manifest measured variable intercepts were constrained to be equal over occasions, we pursued a posteriori models in which there was partial invariance. Using LISREL's modification indices, we evaluated which particular invariance

constraints contributed most substantially to the misfit. Consistent with earlier analyses demonstrating support for the invariance of factor loadings based on the covariance matrix, modification indices associated with factor loadings were consistently smaller than those associated with measured variable intercepts. In the final partially invariant model, invariance constraints were relaxed for three of five measured variable intercepts associated with O5. This suggests that the ease or difficulty in getting high scores apparently shifted at O5; this may not be surprising given that O5 is so temporally distant from the other occasions (the O4/O5 interval is 4 years, whereas other intervals are approximately 1 year). These results provide additional support for interpretations of results for O1 to O4, but may dictate caution in the interpretations of the latent mean associated with O5. Even though the mean difference for O5 is somewhat smaller for the partially invariant model (Model 5A in Table 7), the general conclusions based on this approach are similar to those for the totally invariant mean model (Model 4A in Table 7), but this need not always be so.

More generally, a lack of support for the invariance of measured variable intercepts raises fundamental issues about the interpretation of mean differences. Although we have interpreted the mean differences to reflect shifts in the underlying self-esteem construct, they may also reflect changes in the "difficulty" (i.e., the difficulty in getting a high manifest score on a particular measured variable) of items used to infer self-esteem. If so, the differences across time reflect some unknown combination of changes in "true" self-esteem and changes in the "difficulty" of the items used to infer self-esteem. It is important to emphasize that this problem is not limited to the SEM approach that we have considered here and applies equally to the interpretation of mean scale scores (Table 2) and traditional approaches to evaluating changes in scale scores (e.g., the polynomial contrasts using repeated-measures ANOVA in Table 3). The development of latent mean models is important because such models allow us to evaluate more systematically the implications of these problems.

Relation to Repeated-measures ANOVAs

Beginning with the full moment matrix of multiple indicators, there is a hierarchy of invariance tests and corresponding interpretations that are justified for each level of invariance. In offering this hierarchy of invariances, we emphasize that these absolutely fundamental interpretive issues can only sensibly be discussed or evaluated using multiple-indicator data within a SEM approach. It is therefore useful to evaluate the plausibility of interpretations based on the repeated-measures ANOVA in relation to the structural equation models considered here. In the typical ANOVA approach, critical assumptions include: (a) the complete factorial invariance of factor loadings over time, (b) the complete invariance of manifest intercepts over time, (c)

the homogeneity of uniquenesses and a complete lack of correlated uniquenesses, (d) having multiple indicators that act as parallel forms of the underlying construct, and (e) circularity. If all these assumptions are met, interpretations based on the ANOVA and SEM approaches will be similar and may be justified. Many of these implicit assumptions are not apparent in the ANOVA approach because typically only a single measured variable at each occasion is considered. In fact, the assumptions underlying the ANOVA approach are far more restrictive, less easily tested or completely untestable, and less likely to be met than those in the SEM approach advocated here. Our position is that inferences of genuine latent mean differences on the basis of repeated-measures ANOVAs require a huge leap of faith that can rarely be justified on logical or empirical grounds.

ACKNOWLEDGMENT

The data used in this article were made available by the Inter-University Consortium for Political and Social Research and were originally collected by Jerald Bachman.

REFERENCES

Bachman, J. G. (1970). *Youth in transition: Vol. 2. The impact of family background and intelligence on tenth-grade boys.* Ann Arbor, MI: Institute for Social Research.

Bachman, J. G. (1975). *Youth in transition, Data file documentation: Vol. 2.* Ann Arbor, MI: Institute for Social Research.

Bachman, J. G., & O'Malley, P. M. (1986). Self-concepts, self-esteem, and educational experiences: The frogpond revisited (again). *Journal of Personality and Social Psychology, 50,* 33–46.

Bock, R. D. (1985). *Multivariate statistical methods in behavioral research.* Mooresville, IN: Scientific Software.

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

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

Christofferson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika, 40,* 5–22.

Collins, L. M., & Horn, J. L. (1991). *Best methods for the analysis of change: Recent advances, unanswered questions, future directions.* Washington, DC: American Psychological Association.

Dusek, J. B., Flaherty, J. F. (1981). The development of the self-concept during adolescent years. *Monographs of the Society for Research in Child Development, 46*(4, Whole No. 191).

Goldstein, H. (1979). *The design and analysis of longitudinal studies: Their role in the measurement of change.* London: Academic.

Guttman, L. A. (1954). A new approach to factor analysis: The radex. In P. F. Lazarsfeld (Ed.), *Mathematical thinking in the social sciences* (pp. 258–348). New York: Columbia University Press.

Jöreskog, K. G. (1970). Estimation and testing of simplex models. *British Journal of Mathematical and Statistical Psychology, 23,* 121–145.

Jöreskog, K. G. (1979). Statistical estimation of structural models in longitudinal developmental investigations. In J. R. Nesselroade & P. B. Baltes (Eds.), *Longitudinal research in the study of behavior and development* (pp. 303–352). New York: Academic.

Jöreskog, K. G., & Sörbom, D. (1985). *LISREL VI user's guide.* Mooresville, IN: Scientific Software.

Jöreskog, K. G., & Sörbom, D. (1988). *LISREL VII: Analysis of linear structural relations by the method of maximum likelihood.* Chicago: SPSS.

Markus, H., & Kunda, Z. (1986). Stability and malleability of the self-concept. *Journal of Personality and Social Psychology, 51,* 858–866.

Marsh, H. W. (1987). The big-fish-little-pond effect on academic self-concept. *Journal of Educational Psychology, 79,* 280–295.

Marsh, H. W. (1989). Age and sex effects in multiple dimensions of self-concept: Preadolescence to early-adulthood. *Journal of Educational Psychology, 81,* 417–430.

Marsh, H. W. (1993). Stability of individual differences in multiwave panel studies: Comparison of simplex models and one-factor models. *Journal of Educational Measurement, 30,* 157–183.

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

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indices in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin, 102,* 391–410.

Marsh, H. W., & Grayson, D. (1990). Public/Catholic differences in the high school and beyond data: A multi-group structural equation modelling approach to testing mean differences. *Journal of Educational Statistics, 5,* 199–235.

Marsh, H. W., & Hocevar, D. (1985). The application of confirmatory factor analysis to the study of self-concept: First and higher order structures and their invariance across age groups. *Psychological Bulletin, 97,* 562–582.

McArdle, J. J. (1988). Dynamic but structural equation modeling of repeated measures data. In J. R. Nesselroade & R. B. Catell (Eds.), *Handbook of multivariate experimental psychology* (2nd ed.). New York: Plenum.

McDonald, R. P. (1982). Linear versus nonlinear models in items response theory. *Applied Psychological Measurement, 6,* 379–396.

McDonald, R. P. (1985). *Factor analysis and related methods.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

McDonald, R. P., & Marsh, H. W. (1990). Choosing a multivariate model: Noncentrality and goodness-of-fit. *Psychological Bulletin, 107,* 247–255.

Meredith, W. (1991). Latent variable models for studying differences and change. In L. M. Collins & J. L. Horn (Eds.), *Best methods for the analysis of change: Recent advances, unanswered questions, future directions* (pp. 149–163). Washington, DC: American Psychological Association.

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

O'Malley, P. M., & Bachman, J. G. (1983). Self-esteem: Change and stability between ages 13 and 23. *Developmental Psychology, 19,* 257–268.

Plewis, I. (1985). *Analyzing change: Measurement and explanation using longitudinal data.* New York: Wiley.

Rao, C. R. (1973). *Linear statistical inference and its applications* (3rd ed.). New York: Wiley.

Rogosa, D. R. (1979). Causal models in longitudinal research: Rationale, formulation, and interpretation. In J. R. Nesselroade & P. M. Baltes (Eds.), *Longitudinal research in the study of behavior and development* (pp. 263–302). New York: Academic.

Rosenberg, M. (1985). Self-concept and psychological well-being in adolescence. In R. Leahy (Ed.), *The development of the self* (pp. 205–246). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

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

SPSS, Inc. (1988). *SPSS-X user's guide* (3rd ed.). Chicago: Author.

Walberg, H. J., & Tsaai, S.-L. (1983) Mathews effects in education. *American Educational Research Journal, 20,* 359–373.

Willett, J. B. (1988). Questions and answers in the measurement of change. In E. Z. Rothkopf (Ed.), *Review of research in education* (Vol. 15, pp. 345–422). Washington, DC: American Educational Research Association.

APPENDIX A:
MEANS, STANDARD DEVIATIONS, AND CORRELATIONS AMONG 25 ITEM-PAIRS USED IN THE ANALYSIS

Means

3.8181	3.8272	3.5746	3.7736	3.7903	3.8451	3.8870	3.7173	3.8607	3.8668	3.9631	3.8858	3.7797	3.9239	3.8794
3.9543	3.8702	3.8455	3.9121	3.9326	4.4368	4.1655	4.0875	4.0400	4.3714					

Standard deviations

.7199	.7342	.7671	.6568	.7286	.6890	.6606	.7187	.6093	.6546	.6902	.6799	.7196	.6061	.6587
.6426	.6786	.7031	.6003	.6376	.6194	.6529	.7076	.6190	.5473					

Correlations

1																								
3810	1																							
3500	3994	1																						
3606	3413	4284	1																					
3116	4102	4432	3917	1																				
3915	2489	2391	2254	1742	1																			
2414	3279	2593	2293	2593	4070	1																		
2571	2855	3974	3063	3063	4017	4596	1																	
2334	2389	2100	3269	2502	3903	3813	4600	1																
2123	2476	2734	2433	3382	3033	3755	4731	3544	1															
3673	2647	2360	1970	2371	4884	3245	3309	2911	2301	1														
2244	3007	2477	2269	2222	3187	3629	3960	2796	2875	4625	1													
2298	2482	3505	2355	2695	3001	3348	5395	3290	3410	4490	4988	1												
1894	2019	2284	3158	2475	2840	3151	3898	4264	2959	4374	4548	5006	1											
1797	2175	2544	1978	2797	2499	2885	3948	2925	4268	3219	4696	4986	4066	1										
3005	2629	2328	2163	2202	4209	3017	3265	2696	2448	5108	3667	3961	3068	2942	1									
1609	1903	1916	1459	1848	2665	3282	3096	2463	2400	3097	3549	3636	2953	3624	4396	1								
1780	2097	3404	2268	2214	2618	2759	5042	3150	3424	3550	3721	5347	3851	4452	5009	4954	1							
1850	2208	2561	3025	2451	2446	2678	3669	3621	2831	3042	3094	3985	4415	3463	4461	3982	5493	1						
1475	2260	2967	2336	2773	2007	2762	3946	2625	3507	2656	3327	4012	3527	4548	3730	4299	5683	4536	1					
2133	2003	1910	1646	1710	2884	2259	2596	1643	1665	3477	2433	2652	2143	2006	3873	2459	3037	2308	2384	1				
1155	2273	1661	1291	1678	1852	1996	2584	1905	1718	2075	2472	2490	2218	2182	2762	2534	2780	2309	2504	4623	1			
1056	1709	2078	1676	1600	2090	2166	3388	2071	2339	2261	2653	3770	2299	2886	2767	2833	4075	3220	3295	4835	4713	1		
1350	1798	1877	1618	1393	2234	2177	3010	2727	2226	2347	2252	2898	2907	2560	2893	2385	3029	3368	3059	4465	4594	5601	1	
1044	1436	1312	1477	1920	1198	2299	2477	1586	2896	1885	2286	2528	2505	3182	2371	2272	2941	2814	3123	3366	4160	5109	5108	1

Note. The 25 means and standard deviations are based on responses to 5 item-pairs in Wave 1, 5 item-pairs in Wave 2, and so forth. Correlation matrix presented without decimal points as a lower diagonal matrix.

APPENDIX B:
LISREL 7 SET UP FOR INVARIANT MEANS MODEL WITH POLYNOMIAL CONTRASTS FOR THE LATENT MEANS

```
Stability of latent means with augmented moment matrix
DA  NO = 1,000  NI = 25  MA = am
LA
  "T1P1"  "T1P2"  "T1P3"  "T1P4"  "T1P5"
  "T2P1"  "T2P2"  "T2P3"  "T2P4"  "T2P5"
  "T3P1"  "T3P2"  "T3P3"  "T3P4"  "T3P5"
  "T4P1"  "T4P2"  "T4P3"  "T4P4"  "T4P5"
  "T5P1"  "T5P2"  "T5P3"  "T5P4"  "T5P5"
ME  Fi  = bslfcv2
SD  fi  = bslfcv2
KM  Fi  = bslfcv2  FU RE
MO  NY = 25  NX = 1  NE = 11  FI  LY = FU,FI  PS = SY,FR  TE = SY,FR  BE = FU,FI  GA = FU,FR
PA  PS
0
0  0
0  0  0
0  0  0  0
0  0  0  0  0
0  0  0  0  0  1
0  0  0  0  0  1  1
0  0  0  0  0  1  1  1
0  0  0  0  0  1  1  1  1
0  0  0  0  0  1  1  1  1  1
0  0  0  0  0  0  0  0  0  0  0
PA  LY
1  0  0  0  0  0  0  0  0  0  1
1  0  0  0  0  0  0  0  0  0  1
1  0  0  0  0  0  0  0  0  0  1
```

```
1  0  0  0  0  0  0  0  0  0  1
1  0  0  0  0  0  0  0  0  0  1
0  1  0  0  0  0  0  0  0  0  1
0  1  0  0  0  0  0  0  0  0  1
0  1  0  0  0  0  0  0  0  0  1
0  1  0  0  0  0  0  0  0  0  1
0  1  0  0  0  0  0  0  0  0  1
0  0  1  0  0  0  0  0  0  0  1
0  0  1  0  0  0  0  0  0  0  1
0  0  1  0  0  0  0  0  0  0  1
0  0  1  0  0  0  0  0  0  0  1
0  0  1  0  0  0  0  0  0  0  1
0  0  0  1  0  0  0  0  0  0  1
0  0  0  1  0  0  0  0  0  0  1
0  0  0  1  0  0  0  0  0  0  1
0  0  0  1  0  0  0  0  0  0  1
0  0  0  1  0  0  0  0  0  0  1
0  0  0  0  1  0  0  0  0  0  1
0  0  0  0  1  0  0  0  0  0  1
0  0  0  0  1  0  0  0  0  0  1
0  0  0  0  1  0  0  0  0  0  1
0  0  0  0  1  0  0  0  0  0  1
ST  1  LY(2,1)  LY(7,2)  LY(12,3)  LY(17,4)  LY(22,5)
FI  LY(2,1)  LY(7,2)  LY(12,3)  LY(17,4)  LY(22,5)
PA  TE
1
0  1
0  0  1
0  0  0  1
0  0  0  0  1
1  0  0  0  0  1
```


(Continued)

```
0 1 0 0 0 0 1
0 0 1 0 0 0 0 1
0 0 0 1 0 0 0 0 1
0 0 0 0 1 0 0 0 0 1
1 0 0 0 0 1 0 0 0 0 1
0 1 0 0 0 0 1 0 0 0 0 1
0 0 1 0 0 0 0 1 0 0 0 0 1
0 0 0 1 0 0 0 0 1 0 0 0 0 1
0 0 0 0 1 0 0 0 0 1 0 0 0 0 1
1 0 0 0 0 1 0 0 0 0 1 0 0 0 0 1
0 1 0 0 0 0 1 0 0 0 0 1 0 0 0 0 1
0 0 1 0 0 0 0 1 0 0 0 0 1 0 0 0 0 1
0 0 0 1 0 0 0 0 1 0 0 0 0 1 0 0 0 0 1
0 0 0 0 1 0 0 0 0 1 0 0 0 0 1 0 0 0 0 1
1 0 0 0 0 1 0 0 0 0 1 0 0 0 0 1 0 0 0 0 1
0 1 0 0 0 0 1 0 0 0 0 1 0 0 0 0 1 0 0 0 0 1
0 0 1 0 0 0 0 1 0 0 0 0 1 0 0 0 0 1 0 0 0 0 1
0 0 0 1 0 0 0 0 1 0 0 0 0 1 0 0 0 0 1 0 0 0 0 1
0 0 0 0 1 0 0 0 0 1 0 0 0 0 1 0 0 0 0 1 0 0 0 0 1
eq ly(1,11) ly(6,11) ly(11,11) ly(16,11) ly(21,11)
eq ly(2,11) ly(7,11) ly(12,11) ly(17,11) ly(22,11)
eq ly(3,11) ly(8,11) ly(13,11) ly(18,11) ly(23,11)
eq ly(4,11) ly(9,11) ly(14,11) ly(19,11) ly(24,11)
eq ly(5,11) ly(10,11) ly(15,11) ly(20,11) ly(25,11)
eq ly(1,1) ly(6,2) ly(11,3) ly(16,4) ly(21,5)
eq ly(2,1) ly(7,2) ly(12,3) ly(17,4) ly(22,5)
eq ly(3,1) ly(8,2) ly(13,3) ly(18,4) ly(23,5)
eq ly(4,1) ly(9,2) ly(14,3) ly(19,4) ly(24,5)
eq ly(5,1) ly(10,2) ly(15,3) ly(20,4) ly(25,5)
FI GA(1)-GA(6) GA(11)
```

ma be

0	0	0	0	0	.44721	−.51384	.61679	−.37497	.12222	0
0	0	0	0	0	.44721	−.27759	−.06087	.62779	−.57019	0
0	0	0	0	0	.44721	−.10040	−.38586	.28065	.74984	0
0	0	0	0	0	.44721	.09155	−.56070	−.61617	−.31237	0
0	0	0	0	0	.44721	.80028	.39064	.08270	.01050	0
0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0

MA GA

0	0	0	0	0	0	0.35	0.04	0.03	0.01	1

MA LY

1.1	0	0	0	0	0	0	0	0	0	4.01
1.0	0	0	0	0	0	0	0	0	0	3.92
.97	0	0	0	0	0	0	0	0	0	3.80
.99	0	0	0	0	0	0	0	0	0	3.89
1.1	0	0	0	0	0	0	0	0	0	4.00
0	1.1	0	0	0	0	0	0	0	0	4.01
0	1.0	0	0	0	0	0	0	0	0	3.92
0	.97	0	0	0	0	0	0	0	0	3.80
0	.99	0	0	0	0	0	0	0	0	3.89
0	1.1	0	0	0	0	0	0	0	0	4.00
0	0	1.1	0	0	0	0	0	0	0	4.01
0	0	1.0	0	0	0	0	0	0	0	3.92
0	0	.97	0	0	0	0	0	0	0	3.80
0	0	.99	0	0	0	0	0	0	0	3.89
0	0	1.1	0	0	0	0	0	0	0	4.00
0	0	0	1.1	0	0	0	0	0	0	4.01

(Continued)

APPENDIX B
(Continued)

```
0    0    0    1.0  0    0    0    0    0    0    3.92
0    0    0     .97 0    0    0    0    0    0    3.80
0    0    0     .99 0    0    0    0    0    0    3.89
0    0    0    1.1  0    0    0    0    0    0    4.00
0    0    0    0    1.1  0    0    0    0    0    4.01
0    0    0    0    1.0  0    0    0    0    0    3.92
0    0    0    0     .97 0    0    0    0    0    3.80
0    0    0    0     .99 0    0    0    0    0    3.89
0    0    0    0    1.1  0    0    0    0    0    4.00
MA   PS
0
0    0
0    0    0
0    0    0    0
0    0    0    0    0
0    0    0    0    0    14.24
0    0    0    0    0    14.46  14.82
0    0    0    0    0    14.65  14.98  15.25
0    0    0    0    0    14.71  15.03  15.27  15.39
0    0    0    0    0    15.97  16.32  16.56  16.65  18.17
0    0    0    0    0    0      0      0      0      0      0
MA   TE
 .50
0     .49
0    0     .49
0    0    0     .42
0    0    0    0     .47
 .10 0    0    0    0     .47
```

```
0      .04  0    0    0    0     .42
0    0      .05  0    0    0    0     .41
0    0    0      .05  0    0    0    0     .39
0    0    0    0      .05  0    0    0    0     .43
 .09 0    0    0    0     .11  0    0    0    0     .45
0      .03  0    0    0    0     .02  0    0    0    0     .40
0    0      .05  0    0    0    0     .08  0    0    0    0     .41
0    0    0      .05  0    0    0    0     .06  0    0    0    0     .37
0    0    0    0      .03  0    0    0    0     .07  0    0    0    0     .41
 .06 0    0    0    0     .08  0    0    0    0     .09  0    0    0    0     .40
0    0    0    0    0    0     .03  0    0    0    0     .01  0    0    0    0     .43
0    0      .04  0    0    0    0     .07  0    0    0    0     .06  0    0    0    0     .39
0    0    0      .04  0    0    0    0     .04  0    0    0    0     .04  0    0    0    0     .35
0    0    0    0      .02  0    0    0    0     .04  0    0    0    0     .05  0    0    0    0     .38
 .04 0    0    0    0     .05  0    0    0    0     .07  0    0    0    0     .06  0    0    0    0     .46
0      .03  0    0    0    0     .01  0    0    0    0     .01  0    0    0    0     .01  0    0    0    0     .43
0    0      .02  0    0    0    0     .03  0    0    0    0     .05  0    0    0    0     .04  0    0    0    0     .42
0    0    0      .02  0    0    0    0     .01  0    0    0    0     .01  0    0    0    0     0    0    0    0    0     .37
0    0    0    0      .02  0    0    0    0     .04  0    0    0    0     .03  0    0    0    0     .01  0    0    0    0     .10
OU  ns  SE  TV  MI  SS  ad = off  nd = 3
```

Note. This is the LISREL setup used to estimate polynomial trend components in the latent means model with complete factorial invariance of factor loadings and latent intercepts over time. The means, standard deviations, and correlations for the 25 measured variables used in this analysis are presented in Appendix A. The polynomial contrast coefficients (in the BE matrix) take into account the unequal spacing of the different intervals. Setups involving other sets of contrasts summarized in Tables 5 and 7 differed from the one presented here only in the contrast coefficients that appear in the BE matrix. The setup used to estimate model 4A differed from this in that (a) on the MO line, "BE = FU,FI" is removed and "NE = 6" replaces "NE = 11," (b) the first five columns and rows are removed from the "PA PS" and "MA PS" matrices, (c) the first five zeros are removed from the "MA GA" matrix, (d) Columns 6 through 10 are removed from the "PA LY" and "MA LY" matrices, (e) the five sets of equality constraints of measured-variable intercepts referred to the sixth column of LY instead of the 11th (e.g., "eq ly(1,6) ly(6,6) ly(11,6) ly(16,6) ly(21,6)" instead of "eq ly(1,11) ly(6,11) ly(11,11) ly(16,11) ly(21,11)", and (f) "FI GA(1) – GA(6) GA(11)" is replaced by "FI GA(1) GA(6)." The setup used to estimate the partially invariant Model 5A differed from that for Model 4A only in that the line "FR LY(21,6) LY(25,6) LY(24,6)" was added before the final "OU" line.

APPENDIX C:
LISREL MODEL UNDERLYING THE LISREL 7 SETUP IN APPENDIX B

This Appendix exhibits the mathematical model underlying the LISREL 7 setup shown in Appendix B, and is included as an aid to those wishing to conduct similar analyses. The following model combines a "standard" latent means analysis (Sorbom, 1974) with a transformation of the latent occasion constructs to contrasts (of substantive choice) over these occasions. Such transformations are discussed in the more familiar context of manifest variables in Bock (1985; particularly the addendum titled "Univariate and Multivariate Analysis of Variance of Time-Structured Data").

Consider, firstly, the LISREL model for fitting an augmented raw product–moment matrix (with means) with a structure including latent means and variances. This approach "offers" LISREL a product–moment matrix augmented by a column and row of means (the last diagonal element being unity) and "requests" LISREL to fit it as a covariance matrix, with the following parameters:

$$\begin{pmatrix}\eta\\ \xi\\ 1\end{pmatrix}=\begin{bmatrix}B & \Gamma & 0\\ 0 & 0 & 0\\ 0' & 0' & 0\end{bmatrix}\begin{pmatrix}\eta\\ \xi\\ 1\end{pmatrix}+\begin{bmatrix}\alpha\\ \kappa\\ 1\end{bmatrix}1+\begin{pmatrix}\zeta\\ \xi-\kappa\\ 0\end{pmatrix}$$

$$\begin{pmatrix}y\\ x\end{pmatrix}=\begin{bmatrix}\Lambda_y & 0 & \tau_y\\ 0 & \Lambda_x & \tau_x\end{bmatrix}\begin{pmatrix}\eta\\ \xi\\ 1\end{pmatrix}+\begin{pmatrix}\varepsilon\\ \delta\end{pmatrix}$$

$$1 = 1(\text{FIXED } X)$$

where $E\xi = \kappa$, and ζ, ε, and δ have zero expectations.

If, as in the present "one sample" case, there are no latent exogenous variables and no regressions among the latent constructs, this reduces to:

$$\begin{pmatrix}\eta\\ 1\end{pmatrix}=\begin{bmatrix}B & 0\\ 0' & 0\end{bmatrix}\begin{pmatrix}\eta\\ 1\end{pmatrix}+\begin{bmatrix}\alpha\\ 1\end{bmatrix}1+\begin{pmatrix}\zeta\\ 0\end{pmatrix}$$

$$(y)=\begin{bmatrix}\Lambda_y & \tau_y\end{bmatrix}\begin{pmatrix}\eta\\ 1\end{pmatrix}+(\varepsilon)$$

$$1 = 1(\text{FIXED } X) \qquad (1)$$

where $B = 0$, $E\eta = \alpha$, $\alpha_1 = 0$ for identification and only constructs of the components of α, $\varphi'\alpha\,(\varphi'1 = 0)$, are estimable.

Consider, secondly, any $T \times 1$ random vector η (in general, latent or manifest; but in this article, to be identified with the latent self-esteem constructs on O1, ..., O5). If $E\eta = \alpha$, then $\eta = \alpha + \zeta$, where ζ represents the $T \times 1$ random vector of mean-corrected deviations ($E\zeta = 0$). We can now choose any full rank linear transformation, P, and define new scores $\eta^* = P\eta$. In the present context we are concerned with familiar contrasts, so we restrict attention to P matrices with structure:

$$P=\begin{bmatrix}\varphi_0'\\ \varphi_1'\\ \vdots\\ \varphi_{T-1}'\end{bmatrix}=\begin{bmatrix}\varphi_0(1) & \dots & \varphi_0(T)\\ \varphi_1(1) & \dots & \varphi_1(T)\\ \dots & \dots & \dots\\ \varphi_{T-1}(1) & \dots & \varphi_{T-1}(T)\end{bmatrix},$$

where

$$\varphi_0' = \frac{1}{\sqrt{T}} 1',$$

with

$$\varphi_j' 1 = 0,\ j = 1, \ldots, T-1.$$

The choices discussed in this article for the contrasts φ_j' are trend components (on unequal intervals), difference contrasts and "repeated" contrasts.

The new random vector η^* has mean $\alpha^* = P\alpha$. The last $T-1$ components of α^* give the contrast values for the contrasts selected in the last $T-1$ rows of the matrix P. Mean-correcting the new scores, we obtain the contrast deviation scores (for individual subjects) $\zeta^* = \eta^* - \alpha^*$, and the VC-matrix of the new contrast scores is given by:

$$E\,\zeta^*\,\zeta^{*\prime} = P\Psi P', \text{ where } \Psi = E\,\zeta\,\zeta'.$$

Estimates of the mean and variance of the new contrast scores can be obtained (indirectly) by first fitting the reduced LISREL means model (as in the aforementioned Equation 1, *with* η *being a* $T \times 1$ *vector*), obtaining estimates of α and $\Psi = E\zeta\zeta'$, and then setting $\alpha^* = P\alpha$ and $\Psi^* = P\Psi P'$.

Alternatively, LISREL can be "cajoled" into directly supplying these estimates by interpreting Equation 1 as follows:

$$\begin{pmatrix} \eta \\ \eta^* \\ 1 \end{pmatrix}_{(2T+1)\times 1} = \begin{bmatrix} 0 & P^{-1} & 0 \\ 0 & 0 & 0 \\ 0' & 0' & 0 \end{bmatrix} \begin{pmatrix} \eta \\ \eta^* \\ 1 \end{pmatrix} + \begin{bmatrix} 0 \\ \alpha^* \\ 0 \end{bmatrix} 1 + \begin{pmatrix} 0 \\ \zeta \\ 0 \end{pmatrix}$$

$$(y) = \begin{bmatrix} \Lambda_y & 0 & \tau_y \end{bmatrix} \begin{pmatrix} \eta \\ \eta^* \\ 1 \end{pmatrix} + (\varepsilon)$$

$$1 = 1(\text{FIXED } X)$$

These equations yield $\eta = P^{-1}\eta^*$ and $\eta^* = \alpha^* + \zeta$, consistent with the transformation discussed previously, with $y = \tau_y + \Lambda_y\eta + \varepsilon$ still indicating the "occasion" constructs directly. That is, the terms η, α, Ψ, B, Λ_y, and ζ in LISREL Equation 1 are now, respectively, fully identified with the contrast entities:

$$\begin{pmatrix} \eta \\ \eta^* \end{pmatrix}, \begin{pmatrix} 0 \\ \alpha^* \end{pmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & \Psi^* \end{bmatrix}, \begin{bmatrix} 0 & P^{-1} \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} \Lambda_y & 0 \end{bmatrix}, \begin{pmatrix} 0 \\ \zeta^* \end{pmatrix}.$$

As in Appendix B, *the first component of* α^* *must be set equal to a* constant (0, for example). That is, the first $T + 1$ components in α in the LISREL run based on Equation 1 must be fixed at 0, the rest freely estimating the values of the $T-1$ latent contrasts chosen in P. All components of Ψ ($2T \times 2T$) in the LISREL run are fixed at 0, with the exception of the lower $T \times T$ diagonal block, which freely estimates the VC-matrix of latent variables η^*. Finally, $B(2T \times 2T)$ in the LISREL run is fixed at 0, except for the upper right $T \times T$ block that is fixed at P^{-1}, the inverse of contrasts' matrix. (If and only if the rows of P are chosen to represent orthonormal contrasts, then $P^{-1} = P'$, as with the trend contrasts.)