DOCUMENT RESUME

ED 113 369 95 TM 004 866

AUTHOR Werts, Charles E.; Linn, Robert L.
TITLE Study of Academic Growth Using Simplex Models, Final Report.
INSTITUTION Educational Testing Service, Princeton, N.J.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
BUREAU NO BR-3-0649
PUB DATE Jun 75
GRANT NE-G-00-3-0007
NOTE 66p.

EDRS PRICE MF-$0.76 HC-$3.32 Plus Postage
DESCRIPTORS *Academic Achievement; Achievement Gains; Analysis of Covariance; College Students; Computer Programs; Correlation; Goodness of Fit; Grades (Scholastic); *Longitudinal Studies; *Mathematical Models; Matrices; *Measurement Techniques; Predictor Variables; Scores; Standard Error of Measurement; Statistical Analysis; Testing Problems; *Test Reliability; True Scores
IDENTIFIERS Academic Growth Measurement; Quasi Markov Simplex Model; *Simplex Models; Statistical Variance

ABSTRACT

Forming a sequence covering the various aspects of the simplex model, four articles are presented here under the following titles: "A Simplex Model for Analyzing Academic Growth", "Analyzing Ratings With Correlated Intrajudge Measurement Errors", "The Correlation of States With Gain", and "The Reliability of College Grades from Longitudinal Data". The most important finding of this study is that a simplex model which allows for measurement error, fits a variety of longitudinal academic data quite well. This allows for attenuation corrections when only one measure is available at each time. More importantly, the results suggest that the commonly used split-half or parallel form procedures for estimating reliability may typically yield overestimates or reliability due to "method" variance, i.e., nonindependent measurement errors resulting from the use of closely similar item types. The simplex model appears less subject to this problem because both item format and content change over time. It has been demonstrated that accurate corrections for attenuation are essential to a study of the determinants of academic growth. (Author/RC)

FINAL REPORT

Project No. 3-0649

Grant No. NE-G00-3-0007

STUDY OF ACADEMIC GROWTH USING SIMPLEX MODELS

Charles E. Werts and Robert L. Linn
Educational Testing Service
Princeton, New Jersey 08540

June 1975

The research reported herein was performed pursuant to a grant with the National Institute of Education, U.S. Department of Health, Education, and Welfare. Contractors undertaking such projects under Government sponsorship are encouraged to express freely their professional judgment in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official National Institute of Education position or policy.

U.S. DEPARTMENT OF

HEALTH, EDUCATION, AND WELFARE

National Institute of Education

Table of Contents

Acknowledgments

The authors are indebted to Dr. Karl Jöreskog, Uppsala University. whose ideas provided the basis for our work. Dr. Hubert Blalock, University of Washington, Dr. Arthur Goldberger, University of Wisconsin, and Dr. Frederic Lord, Educational Testing Service, also provided valuable consultation.

A. Objectives

In the research proposal for this project, a variety of tasks were outlined, including:

1. Algebraic formulation of a simplex growth model.

2. Formulation of this model in the format required by Jöreskog's LISREL program.

3. Testing the fit of this simplex model to a variety of longitudinal academic measures.

4. Specifying implications for the study of the determinants of academic growth.

The results for tasks 1 and 2 are given in Section C, those for task 3 in sections C, D, and E, and those for task 4 mostly in section D. All of the original objectives were accomplished.

B. Introduction

Since a project of this type is of little value unless its procedures are made available to other researchers, the work was organized so as to produce publishable journal articles as soon as the relevant results became available. It is these articles which form the basis for this report i.e.,

1. Section C has been accepted by Educational and Psychological Measurement with the title "A Simplex Model for Analyzing Academic Growth".
2. Section F has been accepted by Educational and Psychological Measurement with the title "Analyzing Ratings with Correlated Intrajudge Measurement Errors".
3. Section D has been submitted to Educational and Psychological Measurement as "The Correlation of Status With Gain".
4. Section E has been submitted to Educational and Psychological Measurement as "The Reliability of College Grades from Longitudinal Data".

It was considered desirable to submit all these articles to the same journal mostly because of the appropriateness of its readership, but also because these papers form a sequence which covers the various aspects of the simplex model.

C. Methodology of a Simplex Growth Model

Werts, Jöreskog, and Linn (1972) discuss models for studying growth which require multiple indicators of the growth variable at each measurement period. In this paper a simplex growth model will be considered which needs only a single indicator at each measurement period. Procedures for testing the simplex assumption and for obtaining traditional growth statistics are discussed. The simplex model appears to be particularly appropriate for studies of academic growth (Humphreys, 1960, 1968; Lunneborg, & Lunneborg, 1970). Jöreskog (1970) has provided procedures for the estimation and testing of simplex models. This paper will analyze quasi-Markov simplex models using a more recent estimation procedure provided by Jöreskog and van Thillo (1972) which permits a less complicated and more flexible formulation.

I. The Quasi Markov Simplex

The observed scores (y_i) are assumed to be related to their corresponding true scores (η_i) by the traditional equation:

$$y_i = \eta_i + \varepsilon_i \quad (1)$$

where all ε_i are independent of each other and of all η_i. Jöreskog (1970, sec. 5.6) notes that the simplex structure among the true scores can be stated as

$$\eta_{i+1} = B_i \eta_i + \zeta_{i+1} \quad (2)$$

where all ζ_i are independent and B_i is the true regression weight. As noted by Humphreys (1960), equation (2) implies that the partial correlation between η_i and η_{i+2} is zero with η_{i+1} controlled. The special characteristic of a growth model (Humphreys, 1960) is that successive η_i have the same units of measurement and the difference (Δ_i) between successive η_i is given by:

$$\eta_{i+1} = \eta_i + \Delta_i \quad (3)$$

It follows from equations (2) and (3) that

$$\Delta_i = (B_i - 1)\eta_i + \zeta_{i+1} \quad (4)$$

II. Estimation

In order to estimate the parameters of the above model, a general computer program (LISREL) for estimating a linear structural equation system (Jöreskog, & van Thillo, 1972) was used. The following description is provided by Jöreskog and van Thillo (1972, pp. 2-4):

"Consider random vectors $\eta' = (\eta_1, \eta_2, \ldots, \eta_m)$ and $\xi' = (\xi_1, \xi_2, \ldots, \xi_n)$ of true dependent and independent variables, respectively, and the following system of linear structural relations

$$B\eta = \Gamma\xi + \zeta \tag{5}$$

where $B(m \times m)$ and $\Gamma(m \times n)$ are coefficient matrices and $\zeta' = (\zeta_1, \zeta_2, \ldots, \zeta_m)$ is a random vector of residuals (errors in equations, random disturbance terms). Without loss of generality it may be assumed that $\mathcal{E}(\eta) = \mathcal{E}(\zeta) = 0$ and $\mathcal{E}(\xi) = 0$. It is furthermore assumed that ζ is uncorrelated with ξ and that B is nonsingular.

The vectors η and ξ are not observed but instead vectors $y' = (y_1, y_3, \ldots, y_p)$ and $x' = (x_1, x_2, \ldots, x_q)$ are observed, such that

$$y = \mu + \Lambda_y \eta + \varepsilon \tag{6}$$

$$x = \nu + \Lambda_x \xi + \delta \tag{7}$$

where $\mu = \mathcal{E}(y)$; $\nu = \mathcal{E}(x)$ and ε and δ are vectors of errors of measurement in y and x respectively. The matrices $\Lambda_y(p \times m)$ and $\Lambda_x(q \times n)$ are regression matrices of y on η and of x on ξ, respectively. It is convenient to refer to y and x as the observed variables and η and ξ as the true variables. The errors of measurement are assumed to be uncorrelated with each other and with the true variates.

Let $\Phi(n \times n)$ and $\psi(m \times m)$ be the variance-covariance matrices of ξ and ζ, respectively, Θ_ε^2 and Θ_δ^2 the diagonal matrices of error variances for y and x, respectively. Then it follows, from the above assumptions, that the variance-covariance matrix $\Sigma[(p + q) \times (p + q)]$ of $z = (y', x')'$ is

$$\Sigma = \begin{pmatrix} \Lambda_y (B^{-1}\Gamma\Phi\Gamma' B'^{-1} + B^{-1}\psi B'^{-1})\Lambda_y' + \Theta_\varepsilon^2 & \Lambda_y B^{-1}\Gamma\Phi\Lambda_x' \\ \Lambda_x \Phi \Gamma' B'^{-1} \Lambda_y' & \Lambda_x \Phi \Lambda_x' + \Theta_\delta^2 \end{pmatrix} \tag{8}$$

The elements of Σ are functions of the elements of Λ_y, Λ_x, B, Γ, Φ, ψ, Θ_δ, and Θ_ε. In applications some of these elements are fixed and equal to assigned values. In particular this is so for elements in Λ_y, Λ_x, B and Γ, but we shall allow for fixed values in the other matrices also. For the remaining nonfixed elements of the six parameter matrices one or more subsets may have identical but unknown values. Thus elements in Λ_y, Λ_x, B, Γ, Φ, ψ, Θ_δ, and Θ_ε are of three kinds: (i) <u>fixed parameters</u> that have been assigned given values, (ii) <u>constrained parameters</u> that are unknown but equal to one or more other parameters and (iii) <u>free parameters</u> that are unknown and not constrained to be equal to any other parameter."

Comparison of equations (1) and (2) to the LISREL formulae indicates that for estimation purposes Λ_x, Γ, Φ, and Θ_δ are not required and may be deleted using a program option which specifies no x. Comparison of equations (2) and (6) indicates that Λ_y is an identity matrix. Equation (2) must be changed to $-B_i \eta_i + \eta_{i+1} = \zeta_{i+1}$ to be in equation (5) format with Γ and ξ deleted. The precise form of B will be shown in the example following.

III. Example

For illustrative purposes data reported by Bracht and Hopkins (1972) were analyzed using the simplex model. These data include standard deviations and correlations among the composite achievement scores for eight tests including the Metropolitan Achievement Test (MAT) at grades 1, 2 and 3; the Iowa Tests of Basic Skills (ITBS) at grades 4, 5, 6 and 7; and the Iowa Tests of Educational Development (ITED) at grade 9. Scores are reported in grade-equivalent units for the MAT and ITBS batteries.

In the simplex formulation:

a. The observed scores at each grade level are

$$y' = [y_1, y_2, y_3, y_4, y_5, y_6, y_7, y_8],$$

b. The errors of measurement at each grade are

$$\varepsilon' = [\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4, \varepsilon_5, \varepsilon_6, \varepsilon_7, \varepsilon_8],$$

c. The true scores at each grade are

$$\eta' = [\eta_1, \eta_2, \eta_3, \eta_4, \eta_5, \eta_6, \eta_7, \eta_8],$$

d. The regression residuals among true scores, (specifying $\eta_1 = \zeta_1$) are

$$\zeta' = [\zeta_1, \zeta_2, \zeta_3, \zeta_4, \zeta_5, \zeta_6, \zeta_7, \zeta_8],$$

e. In equation (6) Λ_y is an 8 by 8 identity matrix.

f. In equation (5)

$$\underset{\sim}{B} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -B_1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -B_2 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -B_3 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -B_4 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -B_5 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -B_6 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -B_7 & 1 \end{bmatrix}$$

g. The variance covariance matrix, ψ, of the ζ_i is a diagonal matrix with entries V_{ζ_i} where $i = 1, 2, \ldots 8$.

h. The variances of the $\underset{\sim}{\varepsilon}_i$ are

$$\underset{\sim}{\Theta}_{\varepsilon}^2 = [V_{\varepsilon 1}, V_{\varepsilon 2}, V_{\varepsilon 3}, V_{\varepsilon 4}, V_{\varepsilon 5}, V_{\varepsilon 6}, V_{\varepsilon 7}, V_{\varepsilon 8}].$$

Following Jöreskog (1970) it can be shown that $V_{\varepsilon 1}$, $V_{\varepsilon 8}$, B_1, $V_{\zeta 1}$, $V_{\zeta 2}$ and $V_{\zeta 8}$ are not identified; i.e., unique estimates cannot be obtained. Identification of all parameters was achieved by arbitrarily assigning fixed values to $V_{\varepsilon 1}$ and $V_{\varepsilon 8}$.

Given these additional specifications, there are $8 \times 9 \div 2 = 36$ distinct elements in Σ and 21 parameters to be estimated (6 $V_{\varepsilon i}$, 7 B_i, and 8 $V_{\zeta i}$) which leaves 15 degrees of freedom (overidentifying restrictions) to test the fit of the model to the data.

The observed variance-covariance matrix (S) for the eight variables in the Bracht and Hopkins (1972) data is given in Table 1.

Table 1. Observed variance-covariance matrix

y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8
0.260							
0.257	0.476						
0.336	0.528	0.792					
0.371	0.551	0.782	1.020				
0.416	0.646	0.918	1.127	1.440			
0.437	0.661	0.942	1.158	1.406	1.588		
0.465	0.705	0.995	1.213	1.490	1.634	1.904	
1.576	2.168	3.208	4.005	5.006	5.516	6.112	26.523

LISREL yielded the following maximum-likelihood parameter estimates:

a. Parameters in $\underset{\sim}{B}$: $B_2 = 1.318$, $B_3 = 1.054$, $B_4 = 1.171$, $B_5 = 1.026$, $B_6 = 1.055$ and $B_7 = 3.367$.

b. Parameters in ψ : $V_{\zeta_3} = .049$, $V_{\zeta_4} = .137$, $V_{\zeta_5} = .052$, $V_{\zeta_6} = .107$, $V_{\zeta_7} = .091$.

c. Parameters in $\underset{\sim}{\theta}_{\varepsilon}^2$: $V_{\varepsilon 2} = (.276)^2$, $V_{\varepsilon 3} = (.222)^2$, $V_{\varepsilon 4} = (.240)^2$, $V_{\varepsilon 5} = (.260)^2$, $V_{\varepsilon 6} = (.193)^2$, $V_{\varepsilon 7} = (.298)^2$.

Parameters not listed are not identified. The program also estimates the variance-covariance matrix among the true variables (η_i).

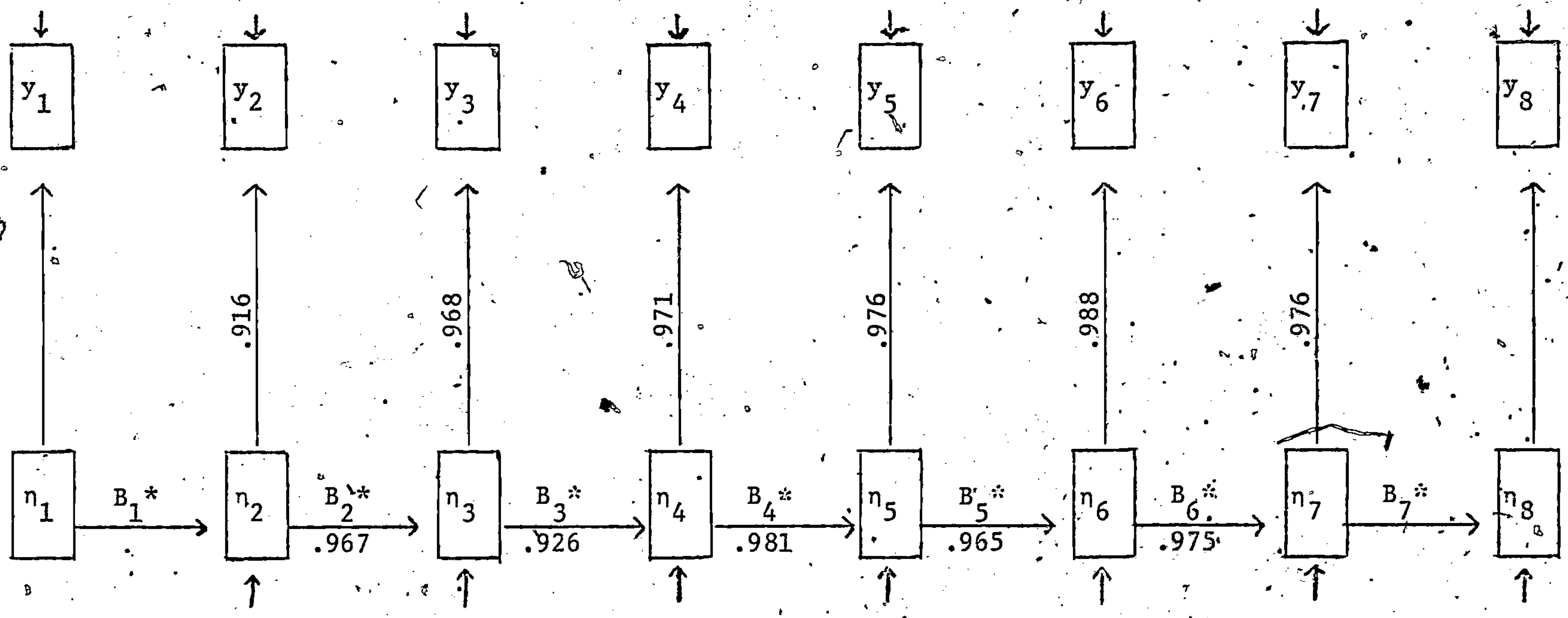
A crucial part of the output is the estimated value of $\hat{\Sigma}$ and the corresponding discrepancies from the observed variance-covariance matrix, S. If these are so large as to indicate an incorrect model, then the above parameter estimates would have little meaning. The residuals of $S - \hat{\Sigma}$ are given in Table 2. Because of the difficulty in comparing residuals between variables with different variances, Table 2 gives discrepancies after S and Σ have been standardized to correlation matrices. Considering the fairly large number of overidentifying restrictions (df = 15) and the fact that S is a missing data matrix with sample sizes ranging from 300 to 1240, these results indicate a reasonably good fit of the model to the data. The goodness of fit test assuming a sample size of 300, yielded a chi-square of 26.3 with 15 degrees of freedom. The probability of getting a chi-squared value larger than that actually obtained, given that the hypothesized model is true is $P = .035$. Because of the relatively large samples this statistic is of minimal interest because quite small discrepancies will be statistically significant.

Table 2. Residuals $S - \Sigma$ standardized

-0.000							
-0.000	0.000						
-0.006	0.002	-0.000					
0.027	-0.007	-0.001	-0.000				
-0.003	-0.005	0.001	0.000	0.000			
0.013	-0.007	0.001	0.001	-0.001	0.000		
0.018	0.002	0.003	-0.005	0.004	-0.000	-0.000	
0.021	-0.056	-0.028	-0.019	0.001	0.002	-0.000	-0.000

To facilitate interpretation of parameter estimates a program option was used which standardizes all parameter estimates and provides the correlation among the true variables. These results are shown in Figure 1. Werts, Jöreskog, & Linn (1971) have shown that the correlation between y_1 and η_2 and between y_8 and η_7 are identified (estimated at .797 and .881 respectively). The estimated reliability of each observed variable is the square of the correlation with the corresponding true score e.g. the reliability of y_2 is $(.916)^2 = .839$. The correlation between any two nonadjacent true variables is equal to the product of the intervening B_i^*. The standardized B_i (B_i^*) are equal to the unattenuated correlations (i.e. corrected for unreliability) between the corresponding variables in the model. The average reliability for y_1 through y_7 is estimated as .93 which compares with .98 reported by Bracht & Hopkins (1972). Our lower estimate might arise (a) because of nonindependent errors of measurement for the various subtests in each composite, resulting in overestimates of reliability by the procedure employed by Bracht & Hopkins and/or (b) because estimates derived from structural models involving theoretical propositions can be expected to reflect both reliability and validity type measurement errors and/or (c) because the model does not fit perfectly.

Figure 1. Simplex Model Estimates Standardized



IV. GROWTH Statistics

A growth model requires variables to have the same units of measurement over time. This is necessary in order for the difference between final and initial status to be meaningful, i.e. subtraction only makes sense when the units are the same. In the example analyzed in section III the scores for the MAT and ITBS batteries were reported in grade-equivalent units. Whether the units are in fact equivalent over time is unknown, however for illustrative purposes they will be so treated.

The variance-covariance matrix for the true factors (η_i) is the basic datum for computation of growth statistics. Werts, Jöreskog, & Linn (1972) have shown that: a. the variance of the gain scores, V_{Δ_i} may be estimated from $\hat{V}_{\Delta_i} = \hat{V}_{\eta_i} + \hat{V}_{\eta_{i+1}} - 2\,\hat{C}(\eta_i\ \eta_{i+1})$ (9),

where $C(\eta_i\ \eta_{i+1})$ is the covariance between η_i and η_{i+1},

b. the true correlation of status with gain,

$\rho(\eta_i\ \Delta_i)$ is given by:

$$\hat{\rho}(\eta_i, \Delta_i) = (\hat{\beta}_i - 1)\sqrt{\hat{V}_{\eta_i} \div \hat{V}_{\Delta_i}} \quad (10), \text{ and}$$

c. the reliability of gain scores, ρ_{Δ_i} is given by:

$$\hat{\rho}_{\Delta_i} = \frac{V_{\Delta_i}}{\hat{V}_{\Delta_i} + V_{\varepsilon_i} + V_{\varepsilon_{i+1}}} \quad (11).$$

The estimated variance-covariance matrix among the true variables is given in table 3 except for the unidentified variances of η_1 and η_8. For this reason no growth statistics involving η_1 and η_8 can be computed. Comparison of tables 1 and 3 will show that the covariances between the true variables approximate those between the corresponding observed variables. If the model is correct any discrepancies would be ascribed to sampling errors since according to equation (1) the observed and true covariances should be identical.

Table 3. Variance - covariance Matrix Among True Variables

η_1	η_2	η_3	η_4	η_5	η_6	η_7	η_8
*							
0.257	0.400						
0.338	0.527	0.743					
0.357	0.555	0.783	0.962				
0.418	0.650	0.917	1.127	1.372			
0.428	0.667	0.940	1.156	1.407	1.550		
0.452	0.703	0.992	1.219	1.484	1.635	1.815	
1.521	2.368	3.339	4.104	4.997	5.505	6.112	*

*Not identified

The results in Table 3 were used in equations (9), (10) and (11) to compute the growth statistics presented in Table 4. The average true correlation of status with gain is estimated to be .39 (Fisher's Z transformation used for averaging). The average reliability of gain scores is estimated to be .46; i.e., the order of magnitude of the true change variance approximates that of the associated errors of measurement (θ^2). Table 3 could be used with equations (9), (10), and (11) to compute change statistics between any two true variables e.g. the $\eta_7 - \eta_2$ period yields a change variance of .809, a correlation of status with gain of .533 and a gain reliability of .831. The meaningfulness of these statistics is dependent on the correctness of the assumption of equivalent units of measurement over time.

Table 4. Estimated Growth Statistics

Change	Change Variance	$B_i - 1$	Status - Gain Correlation	Reliability of Gain
$\eta_3 - \eta_2$	.089	.318	.674	.415
$\eta_4 - \eta_3$	.139	.054	.125	.565
$\eta_5 - \eta_4$	.081	.171	.589	.393
$\eta_6 - \eta_5$	.108	.026	.093	.507
$\eta_7 - \eta_6$	.096	.055	.221	.432

V. A Structural Model for Growth

The above estimation model used only the simplex equations (1) and (2). For estimation purposes equations (3) and (4) could have been used directly to define the model. It follows from equation (3) that $\eta_{i+1} = \eta_1 + \sum_{1}^{i} \Delta_i$ in which case the vector of true variables becomes:

$$\eta' = (\eta_1, \Delta_1, \Delta_2, \Delta_3, \Delta_4, \Delta_5, \Delta_6, \Delta_7) \text{ and}$$

y', ε', and ζ' remain the same as before. Equation (1) becomes:

$$y_i = (\eta_1 + \sum_{1}^{i} \Delta_i) + \varepsilon_i \qquad (12)$$

and equation (4):

$$\Delta_i = (B_i - 1)(\eta_i + \sum_{1}^{i} \Delta_i) + \zeta_{i+1} \qquad (13)$$

Translating these equations into equations (5) and (6):

Θ^2_ε and ψ remain the same. It follows from equation (12) in the example:

$$\Lambda_y \Delta = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

Translating equation (13) and defining $b_i = (B_i - 1)$:

$$B_\Delta = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ b_1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ b_2 & b_2 & 1 & 0 & 0 & 0 & 0 & 0 \\ b_3 & b_3 & b_3 & 1 & 0 & 0 & 0 & 0 \\ b_4 & b_4 & b_4 & b_4 & 1 & 0 & 0 & 0 \\ b_5 & b_5 & b_5 & b_5 & b_5 & 1 & 0 & 0 \\ b_6 & b_6 & b_6 & b_6 & b_6 & b_6 & 1 & 0 \\ b_7 & b_7 & b_7 & b_7 & b_7 & b_7 & b_7 & 1 \end{bmatrix}$$

This model is simply a linear transformation of the previous model, has the same number of parameters to be estimated (V_{ε_1} & V_{ε_8} arbitrarily fixed), and the model should (and did) fit the data to the same degree as in the previous analysis. It should be noted however that this formulation required use (in B_Δ) of the LISREL option permitting parameters to be specified as equal.

The variance-covariance matrix of the true variables ($\hat{C}$) in the growth formulation yields the estimated true change variance ($\hat{V}_{\Delta_m}$) directly. It is reasonable to use this formulation in conjunction with the previous formulation to obtain the change statistics. Also of supplemental interest is the standardized $\hat{C}$ which gives the correlation among the change factors as shown in Table 5 (Δ_1 and Δ_7 are not identified). It can be seen that most of the correlations are low positive. Differences between these correlations are difficult to interpret because of fluctuation in the associated change variances.

Table 5. Correlations Among True Change Scores

	Δ_2	Δ_3	Δ_4	Δ_5	Δ_6
Δ_2	1.000				
Δ_3	.104	1.000			
Δ_4	.460	.290	1.000		
Δ_5	.069	.044	.067	1.000	
Δ_6	.162	.102	.156	.077	1.000

VI. Estimating True Growth

Werts, Jöreskog and Linn (1972) give a procedure for estimating true growth scores from all observed measures in a structural model. The basic problem is to obtain a variance-covariance matrix for the observed variables and the true change to be predicted e.g. take $\Delta_4 = \eta_5 - \eta_4$ as the variable to be predicted. It follows from equations (1), (2), (3) and (4) that the covariance between the y_i and Δ_4 are given by:

$$C(y_1\ \Delta_4) = [B_4 - 1]\ B_3\ B_2\ C(\eta_1\eta_2)$$

$$C(y_2\ \Delta_4) = [B_4 - 1]\ B_3\ B_2\ V_{\eta_2}$$

$$C(y_3\ \Delta_4) = [B_4 - 1]\ B_3\ V_{\eta_3}$$

$$C(y_4\ \Delta_4) = [B_4 - 1]\ V_{\eta_4}$$

$$C(y_5\ \Delta_4) = [B_4 - 1]\ V_{\eta_4} + V_{\Delta_4}$$

$$C(y_6\ \Delta_4) = B_5\ C(y_5\ \Delta_4)$$

$$C(y_7\ \Delta_4) = B_6\ C(y_6\ \Delta_4)$$

$$C(y_8\ \Delta_4) = B_7\ C(y_7\ \Delta_4)$$

For the purpose of estimating Δ_4 it is appropriate to use the estimated variances and covariances among the y_i as obtained from $\hat{\Sigma}$ rather than the observed matrix because $\hat{\Sigma}$ is presumed to be the best population estimate of these values when the model is accepted. The resulting variance-covariance matrix is given in table 6 in standardized form to facilitate interpretation. LISREL was then set up for the regression of Δ_4 on the y_i. The y_i predicted 63% of the variance in Δ_4. This is not much larger than the 59% of the

Table 6. Correlations Among y_i and Δ_4

	Δ_4	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8
Δ_4	1.000								
y_1	.421	1.000							
y_2	.484	.731	1.000						
y_3	.529	.747	.858	1.000					
y_4	.572	.693	.797	.871	1.000				
y_5	.719	.683	.785	.859	.930	1.000			
y_6	.702	.666	.767	.838	.908	.930	1.000		
y_7	.677	.642	.738	.808	.875	.896	.940	1.000	
y_8	.610	.579	.666	.729	.789	.809	.848	.860	1.000

variance predictable from y_4 and y_5 alone because the reliabilities of y_4 and y_5 are quite high. Note that the Δ_4 reliability of .39 is not directly comparable to these figures because reliability corresponds to the proportion of observed variance predictable from the true score whereas our procedure is the reverse.

VII. Discussion

In order to understand the value of the quasi Markov simplex model for studies of academic growth it is useful to detail the rationale behind the estimation of the reliabilities of each time. For example consider the reliability for y_5 i.e. the squared correlation of y_5 and η_5.

Given one measure prior to the fifth grade (m = 1,2,3,4) and one measure subsequent to the fifth grade (n = 5,7,8) then in the simplex model the reliability of y_5 is

$$\rho(y_5\ \eta_5)^2 = \frac{\rho(y_m\ y_5)\ \rho(y_5\ y_n)}{\rho(y_m\ y_n)}$$

When the model fits the data the implication is that except for sampling variations the estimates of $\rho(y_5\ \eta_5)$ derived from the various combinations of y_m and y_n will be equal. The greater the number of consistent estimates (m x n = 12), the more generalizable the results are likely to be. This method contrasts with the split half or parallel form methods of obtaining test reliabilities which provide a single estimate which cannot be rejected because of inconsistency with the data (i.e. it is "just" identified). The split half or parallel form procedures involve almost identical item formats which could well lead to overestimation of reliability because of the presence of "method" variance. The simplex model approach is less subject to method variance because over time both item format and content change. This is probably the main reason that in the example the reliabilities from the simplex model were less than that reported in the data source.

Bibliography

Bracht, G. H. & Hopkins, K. D. Stability of educational achievement. In Bracht, G. H., Hopkins, K. D. & Stanley, J. C. (eds.) Perspectives in Educational Measurement, Prentice-Hall, Englewood Cliffs, N. J., 1972

Humphreys, L. G. Investigations of the simplex. Psychometrica, 1960, 25, 313-323.

Humphreys, L. G. The fleeting nature of college academic success. Journal of Educational Psychology, 1968, 59, 375-380.

Jöreskog, K. G. Estimation and testing of simplex models. The British Journal of Mathematical and Statistical Psychology, 1970, 23, 121-145.

Jöreskog, K. G. & van Thillo, M. LISREL: A general computer program for estimating a linear structural equation system involving multiple indicators of unmeasured variables. RB-72-56, Princeton, N. J.: Educational Testing Service, 1972.

Lunneborg, C. E. & Lunneborg, P. W. Relations between aptitude changes and academic success during college. Journal of Educational Psychology, 1970, 61, 169-173.

Werts, C. E., Jöreskog, K. G. & Linn, R. L. Comments on "The estimation of measurement error in panel data." American Sociological Review, 1971, 36, 110-113.

Werts, C. E., Jöreskog, K. G., and Linn, R. L. A multitrait-multimethod model for studying growth. Educational and Psychological Measurement, 1972, 32, 655-678.

D. Implications for the Study of Growth

Thorndike (1966) notes that in order to accurately estimate the correlation between initial intellectual status and subsequent intellectual growth it is necessary to have measures expressed in meaningfully equal units which at all ages refer to identically the same attribute of the individual and to either have error-free measures or accurate reliabilities. Because error-free measures are not available corrections are typically made using available reliability coefficients. However, as Lord (1963) has noted, the need to make corrections for attenuation "....poses somewhat of a dilemma, since, first, it is often hard to obtain the particular kind of reliability coefficients that are required for making the appropriate correction," and the correction"... may be seriously affected by sampling errors". Because of the fragility of gain scores it is particularly important to have an accurate reliability estimate. As pointed out by Campbell and Fiske (1959), procedures based on similar measurement methods (e.g., the usual internal consistency or parallel form reliabilities) will be biased because of method variance (i.e. correlated measurement error). In this paper unreliability estimates will be generated using a procedure less subject to method variance which is based on the simplex model (Humphreys, 1960).

Method

The analytical procedures used in this paper are detailed in Werts, Linn, and Jöreskog (in press). A general computer program for estimating a linear structural equation system involving multiple indicators of unmeasured variables called LISREL (Jöreskog and van Thillo, 1972) was used for all computations.

The basic model used in the analyses is called a "quasi-Markov simplex" by Jöreskog (1970). Each observed test score (X_i) is assumed to consist of a true component (T_i) and an independent error (E_i):

$$X_i = T_i + E_i \qquad (1)$$

All of the E_i are assumed independent and successive T_i are related by the linear regression equation:

$$T_{i+1} = B_i T_i + d_i, \qquad (2)$$

where the d_i residuals are independent. The reliability (R_{ii}) of an observed X_i is equal to $(n<i, m>i)$:

$$R_{ii} = (R_{in} \; R_{im}) \div R_{nm} \qquad (3).$$

where R_{in} is the correlation of X_i and a prior measure X_n, R_{im} is the correlation of X_i and a later measure X_m, and R_{nm} is the correlation of X_n and X_m. If there is more than one observed score prior to or following X_i, then there will be more than one possible estimate of R_{ii}. If the simplex model fits the data, then all possible estimates of R_{ii} will be equal within the limits of sampling error. It follows from equation (3) that reliabilities cannot be estimated for either the first or the last measures. When successive measures are on the same scale, then changes in successive true scores (Δ_i) over time, are defined by: $\Delta_i = T_{i+1} - T_i$. For analytical purposes, estimates of the variances (V_{T_i}) of the T_i, the true regression weights (β_i), the reliabilities (R_{ii}), the true change variances ($V_{\Delta i}$), and the true correlation of status with gain ($\rho_{T_i\Delta i}$) will be most relevant.

Data

The data for this study were collected in a ten-year longitudinal study of academic growth at Educational Testing Service (Hilton, Beaton, and Bower 1971).

The School and College Ability Test (SCAT) and Sequential Test of Educational Progress (STEP) were given in the fifth, seventh, ninth, and eleventh grades. SCAT was designed to measure basic verbal and quantitative abilities and provides Verbal, Quantitative and Total scores. STEP was designed to measure skills and problem-solving abilities which are generally considered major goals of educations and yields six subtest scores: Mathematics, Science, Social Studies, Reading, Listening and Writing. The analysis was done on the 2,483 students for whom SCAT and STEP scores were available for all four grades. The variance-covariance matrices for these test scores is given in Table 1 in which the first four columns give the variances for the four occasions and the next six columns give covariances between these occasions.

TEST	V_1	V_2	V_3	V_4	C_{12}	C_{13}	C_{14}	C_{23}	C_{24}	C_{34}
STEP:										
Math	109.992	179.431	184.520	251.172	99.758	100.508	105.562	136.468	142.448	154.895
Science	150.587	121.895	185.270	160.501	97.724	116.270	98.472	110.890	90.232	121.433
Social S.	130.444	172.867	222.648	216.681	117.774	126.673	117.584	154.653	146.644	168.819
Reading	252.143	299.958	256.291	298.194	210.990	183.106	195.151	216.601	222.691	212.424
Listening	150.438	195.429	214.374	216.652	132.970	130.916	120.145	161.556	146.774	156.870
Writing	198.500	261.488	313.265	320.804	173.126	180.092	175.054	222.842	213.632	239.027
SCAT:										
Verbal	135.320	157.242	183.920	208.363	124.718	127.706	132.553	149.294	153.476	171.799
Quant.	73.000	164.858	265.328	310.609	80.160	96.224	100.512	157.130	165.122	231.356
Total	58.722	104.044	157.417	199.128	66.520	78.259	85.104	112.751	122.638	159.112

Table 1. Observed Variance-Covariance Matrices

Testing the Simplex Fit For Each Test

The most crucial part of the analysis concerns the fit of the simplex model to the data. If the model is inconsistent with the data, then parameters estimates are meaningless. Jöreskog's LISREL program provides a large sample chi-square statistic for testing the fit of the model to the data. In essence, this chi-square is a measure of how close the "reproduced" matrix is to the observed varianced-covariance matrix. The "reproduced" matrix is the matrix generated by the estimated maximum likelihood parameter estimates. When sample sizes are large, as in this case, quite small differences between the observed and reproduced matrices will be statistically significant. To help judge the meaningfulness of these differences, both matrices were converted to correlations and the root mean square of the differences between the observed and reproduced correlation matrices was calculated.

In Table 2, the chi-square is given for each of the tests. The Science and Reading subtests show a statistically significant lack of fit as seen from the significance levels given in the second column. However, the root mean square difference between the corresponding observed and reproduced correlation matrices is only .005 and .004 respectively. Such small differences are clearly not meaningful. We conclude that the simplex model provides an excellent fit to the observed data for the four occasions.

TEST	BASIC χ^2	BASIC P	RMS Residuals
STEP:			
Math	0.17	.683	.001
Science	5.39*	.020	.005
Social S.	3.09	.079	.003
Reading	8.06*	.005	.004
Listening	0.53	.468	.001
Writing	1.06	.303	.002
SCAT:			
Verbal	2.18	.140	.001
Quant.	0.24	.623	.001
Total	0.00	.973	.000

Table 2. Chi-square Goodness of Fit Tests

*Significant at the .02 level or better when the χ^2 is tested with one degree of freedom.

Estimated Growth Statistics

Using formulae from Werts, Linn, and Jöreskog (in press) various growth statistics were estimated. Since the reliabilities for the fifth and eleventh grades are not determined by the model, it follows that V_{T1}, V_{T4}, B_1, R_{11}, R_{44}, $V_{\Delta 1}$, V_3, $\rho_{T1\Delta 1}$, $\rho_{T3\Delta 3}$ cannot be estimated.

Parameter estimates for each test are given in Table 3. It can be seen that the true variance increases from the seventh to ninth grades (i.e., V_{T2} to V_{T3}) for all tests except Reading. In interpreting the seventh to ninth (B_2) and ninth to eleventh (B_3) grade true regression weights, it should be noted that a weight of 1.0 means a zero correlation of true status with true gain, a weight greater than 1.0 means a positive correlation and a weight less than 1.0 means a negative correlation. B_2 and B_3 were tested to see if they were significantly different from 1.0 (Werts, Linn & Jöreskog in press). These results (Table 3) indicate that B_2 is less than 1.0 only for Reading and B_3 is less than 1.0 for Science, Social Studies, Listening and Writing. The significance test for $B_2 = 1.0$ is the significance level for $\rho_{T2\Delta 2}$ in the last column of Table 3. The true change variance from seventh to ninth grades $V_{\Delta 2}$ can be compared to the sum of the error variances for the tests at these times (labelled $V_{\Delta E}$ in Table 3).

The reliability of these gain scores would be $V_{\Delta 2} \div V_{\Delta 2} + V_{\Delta E}$. It can be observed that the true variance is quite small compared to the error variance for the STEP subtests but is much more comparable for the SCAT subtests. This is a function of reliabilities and merely points out that obtaining accurate change statistics is possible only with highly reliable tests.

TEST	VT_2	VT_3	B_2	B_3	R_{22}	R_{33}	$V_{\Delta 2}$	$V_{\Delta E}$	$\rho_{T2\Delta 2}$
STEP									
Math	135.000	147.771	1.010	1.047*	.753	.802	10.097	81.032	+.037
Science	92.069	146.752	1.198*	0.827*	.755	.792	18.167	68.167	+.446
Social S.	145.015	179.352	1.070*	0.941*	.839	.806	14.151	71.154	+.224
Reading	245.792	203.956	0.876*	1.042*	.819	.796	18.912	106.515	-.447
Listening	163.556	172.089	0.987	0.912*	.837	.803	12.883	74.417	-.046
Writing	213.038	248.105	1.044*	0.963*	.815	.792	16.337	113.613	+.159
SCAT:									
Verbal	145.285	166.737	1.027*	1.030*	.924	.907	13.684	29.139	+.088
Quant.	132.069	221.339	1.197*	1.049*	.796	.831	37.352	78.238	+.370
Total	95.846	146.293	1.176*	1.088*	.921	.929	16.631	19.319	+.423

Table 3. Maximum Likelihood Parameter Estimates

*Statistically different from one at the .03 level of significance. The chi-square for testing $B_2 = 1$ were .21, 75.07, 15.46, 56.56, 0.61, 5.81, 4.90, 78.17, and 170.62 with significance levels of .33, .00, .00, .00, .55, .02, .03, .00, and .00 respectively. The chi-square for testing $B_3 = 1$ were 4.85, 81.06, 12.43, 5.46, 23.43, 4.32, 6.78, 8.38, and 62.73 with approximate significance levels of .03, .00, .00, .02, .00, .03, .01, and .01 respectively

Methodological Considerations

As noted above the true change variance is typically small compared to the observed variances or the estimated error variances. This means that accurate corrections for attenuation are essential since a small or moderate error in estimating unreliability will normally result in a relatively large error in estimating true change variance. The usual procedures for estimating reliability involve split half or parallel form methods which involve almost identical item formats which could well lead to overestimation of reliability because some of the item covariance is due to "method" variance (Campbell and Fiske, 1969). The simplex model approach is less subject to method variance because over time both item format and content change. This is probably the main reason that reliabilities from the simplex model are usually less than those reported by the test makers. It is not unlikely that many studies of the determinants of academic growth (or change) failed to find correlates of change because of inadequate corrections for unreliability.

The first crucial step in any study of school effects is to measure the changes or growth in cognitive skills during the period of interest. In other words it is necessary to know precisely a person's skills at the start and at the end in order to specify what was learned during the period. Thorndike (1966) pointed out that this requires having the initial and final measures in meaningfully equal units which refer to identically the same attribute of the individual. The logic of this requirement is simply that if the final score is 7 pears and the initial score is 4 apples we can neither specify the number of pears nor the number of apples gained during the period. For example, if the final test measures reasoning ability and the initial test rote memory it will be impossible to know how much either ability has progressed in the interim. If the final score is 7 large apples and the initial score 3 small apples, the change is at least 4 large apples. However, if the final score units were small apples and the initial score units large apples, the gain is difficult to specify. Thus, even if accurate corrections for attenuation are possible, growth may be easily obscured by problems of scaling the units of measurement over time.

The results in Table 3 should make it clear that the observed correlations can have a simplex pattern when the true correlation of status with gain is positive, zero, or negative. Assuming independent errors, if the true gain is uncorrelated with true initial status, then the observed correlations will have a simplex pattern. It does not follow, however, that if a simplex correlational pattern is observed that the correlation of status with gain is zero as has been suggested by Humphreys (1960), Andersen, (1939), and Bloom (1964). Furthermore, the results in Table 3 suggest that there may not be a single true correlation between intellectual status and intellectual growth. The status gain correlation of +.09 for SCAT Verbal and +.37 for SCAT Quantitative based on quite reliable tests, suggest that learning quantitative intellectual skills may be more dependent on prior learning than verbal skills.

Anderson, J.E. The limitation of infant and pre-school tests in the measurement of intelligence. Journal of Psychology, 1939, 8, 351-379.

Bloom, B.S. Stability and change in human characteristics. New York: Wiley, 1964.

Campbell, D.T., Fiske, D.W. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 1959, 56, 81-105.

Hilton, T.L., Beaton, A.E., and Bower, C.P. Stability and instability in academic growth - a compilation of longitudinal data. Final Report, United States Office of Education, Research No. 0-0140, August, 1971. (ERIC number EDO 72075).

Humphreys, L.G. Investigations of the simplex. Psychometrica, 1960, 25, 313-323

Jöreskog, K.G. Estimation and testing of simplex models. The British Journal of Mathematical and Statistical Psychology, 1970, 23, 121-145.

Jöreskog, K.G. and van Thillo, M. LISREL: A general computer program for estimating a linear structural equation system involving multiple indicators of unmeasured variables. RB-72-56. Princeton, N.J.: Educational Testing Service, 1972.

Lord, F.M. Elementary models for measuring change. In C.W. Harris (Ed.), Problems in measuring change. Madison, Wisc.: University of Wisconsin Press, 1963. Pp 21-38

Thorndike, R.L. Intellectual status and intellectual growth. Journal of Educational Psychology, 1966, 57, 121-127.

Werts, C.E., Linn, R.L., and Jöreskog, K.G. A simplex model for analyzing academic growth. Educational and Psychological Measurement, in press.

E. Analyses of Longitudinal Grade Data

Humphreys (1968) notes that the correlations among eight semesters of undergraduate grade-point averages have a simplex form. By this he means that the farther apart the averages are in time the lower will be the correlation between them. Furthermore, Humphreys (1960) notes: "If one is sufficiently confident that the variables do form a simplex, a reliability estimate can be obtained from the intercorrelations of the variables." In this paper a procedure developed by Jöreskog (1970a) will be used to test whether Humphreys' (1968) college grade data form a simplex and to obtain estimates of reliabilities and unattenuated correlations between grades.

I. The Model

The model used by Humphreys is called a "quasi-Markov simplex" by Jöreskog (1970a). In this model each observed grade score (X_i) is composed of a true component (T_i) and an independent error (ε_i) of measurement:

$$X_i = T_i + \varepsilon_i \quad . \tag{1}$$

All of the ε_i are assumed independent and successive T_i are related by the linear regression equation:

$$T_{i+1} = B_i T_i + d_i \quad , \tag{2}$$

where the d_i residuals are assumed independent of each other.

In this model the reliability (r_{ii}) of an observed X_i is equal to $(n<i, m>i)$:

$$r_{ii} = \frac{r_{in}\, r_{im}}{r_{nm}} \tag{3}$$

where r_{in} is the correlation of X_i and X_n, r_{im} is the correlation of X_i and X_m and r_{nm} is the correlation of X_n and X_m. If there is more than one observed score prior to or following X_i, then there will be more than one possible estimate of r_{ii}. If the simplex model fits the data then all the possible estimates of r_{ii} will be equal within the limits of sampling error. It follows from equation (3) that reliabilities cannot be estimated for either the first or the last observed measures.

Jöreskog's (1970a) procedures for the estimation and testing of simplex models was used. This method provides a chi square goodness of fit test and also shows how well the estimated parameters reproduce the observed correlation matrix. The details of this analytical procedure are beyond the scope of this paper.

II. Analysis

The correlation matrix shown in Table 1 was obtained from Humphreys (1968, Table 2). The variables include eight semesters of grade-point averages, high school rank, and composite score on the American College Testing program tests for approximately 1,600 students at the University of Illinois.

A. Simplex Model for Eight Semesters Grade-Point Averages

The initial analysis was designed to test the fit of the simplex model to the correlations among the eight semesters grades. The goodness of fit test yielded a chi square of 23.91 with 15 degrees of freedom. The probability of getting a chi-squared value larger than that actually obtained,

Table 1

Correlations among Observed Variables

	X_0	X'_0	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
X_0	1.000									
X'_0	.393	1.000								
X_1	.387	.375	1.000							
X_2	.341	.298	.556	1.000						
X_3	.278	.237	.456	.490	1.000					
X_4	.270	.255	.439	.445	.562	1.000				
X_5	.240	.238	.415	.418	.496	.512	1.000			
X_6	.256	.252	.399	.383	.456	.469	.551	1.000		
X_7	.240	.219	.387	.364	.445	.442	.500	.544	1.000	
X_8	.222	.173	.342	.339	.354	.416	.453	.482	.541	1.000

Note: X_0 is high school rank, X'_0 ACT composite score, and X_1 through X_8 are eight semesters grade-point averages.

given that the hypothesized model is true, is $P = 0.07$. Especially considering the large sample size, these results are consistent with the hypothesis that the simplex model provides a good fit to the data. If the simplex model is rejected, estimates of model parameters would not be relevant. The estimated parameters shown in Figure 1, include correlations between X_i and T_i and the correlations among the T_i. Reliabilities are equal to the square of the correlation between the corresponding X_i and T_i; e.g., $r_{22} = (.754)^2 = .569$. Although not shown, the correlation of X_1 with T_2 is .737 and of X_8 with T_7 is .694. The maximum likelihood estimates given in Figure 1 could be used to generate the correlations among the observed variables; e.g., $r_{23} = (.754)(.838)(.758) = .479$. The estimated correlations generated in this manner differ from the observed correlations only because of sampling errors, if the model is correct. The estimated correlations are therefore estimated population values given the simplex model. In Table 2 the discrepancies between the observed and the estimated correlations are shown. The small size of these discrepancies is consistent with the chi square statistic in suggesting a good fit of the model to the data. Unlike the chi square, however, the discrepancies do not increase as a function of the sample size.

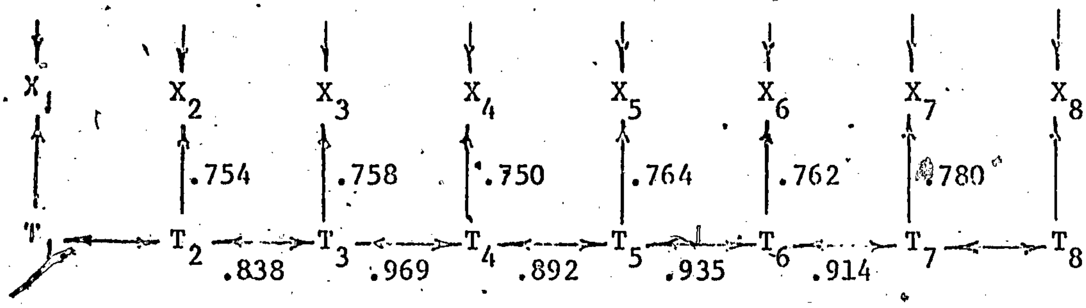


Figure 1. Simplex parameter estimates

Table 2

Discrepancies between Observed and Estimated Correlations among Observed Variables

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
X_1	-.000							
X_2	-.002	.004						
X_3	-.011	.008	.002					
X_4	-.009	-.018	.008	.008				
X_5	.012	.002	-.002	-.002	-.002			
X_6	.023	-.006	-.010	-.011	.007	-.001		
X_7	.037	.003	.012	-.004	-.005	.005	-.010	
X_8	.025	.012	-.038	.012	-.005	.006	.007	-.000

B. Assumption of Equal Reliabilities

The data led Humphreys to believe that the reliabilities across semesters were equal. Jöreskog's (1970a) procedures allow this hypothesis to be tested because model parameters may be constrained to be equal. Because of special features of Jöreskog's simplex analysis the reliabilities could not be directly constrained, however the same effect was obtained by setting the error variances equal. As a result of these constraints the chi square increased to 26.08 with 20 degrees of freedom. The increase in chi square of 2.17 (i.e., 26.08-23.91) with 5 degrees of freedom (i.e., 20-15) is an appropriate test of the equal reliabilities hypothesis. Since the probability of obtaining a larger chi square is approximately .80 this hypothesis was not rejected. These results therefore support Humphreys' conclusion that the reliabilities are equal. The estimated reliability is .579 which corresponds to a correlation of .761 between X_i and T_i in Figure 1. Reading from left to right in Figure 1, the new true correlations, assuming equal reliability, are .836, .965, .891, .936, and .922. These differ very little from those in Figure 1. If it is assumed that X_1 and X_7 have a reliability of .579 then the correlations of T_1 and T_2 would be .966 and that of T_7 and T_8 .914.

C. Modelling

The computer program used for this analysis may be used for a wide variety of other structural models (Jöreskog, 1970b), some of which might include the simplex as a component part. To illustrate this, a model will be hypothesized which includes high school rank (X_0) and the ACT composite score (X_0'). It seems reasonable to suppose that both these variables are measures of high school achievement, i.e.,

$X_0 = T_0 + \varepsilon_0$ and

$X_0' = T_0 + \varepsilon_0'$.

If high school achievement also is part of a simplex pattern then T_0 can be included in equation (2). As before, all errors (including ε_0 and ε_0') are assumed independent. Building on prior results it will be assumed that the college grades have equal reliabilities. A special feature of this model is that the reliability of X_1 can be estimated from equation (3) using either X_0 or X_0' as a prior variable. The analysis of this model yielded a chi square of 45.22 with 34 degrees of freedom. Since the probability of a larger chi square is .095, these results suggest that the hypothesized model is consistent with the data. The discrepancies between observed and estimated correlations are on the order of those given in Table 2. The estimated parameters are given in Figure 2. The reliability of college grade-point averages is estimated as .584 which does not meaningfully differ from previous results. Comparable correlations among true scores also differ little from Figure 1. The estimated reliability of high school rank is .424 and that of the ACT composite .365. The ACT composite reliability is substantially lower than would be expected if parallel form or internal consistency estimates were obtained in which case a value closer to .9

might be expected. The reason for this low value may be that ACT composite was the only test score included in the model. Thus, there may be substantial systematic variability in ε_o' but it is simply uncorrelated with the grade true scores. If two ACT composite scores were available for each student we might postulate a model such as the one shown in Figure 3. According to the model in Figure 3 the observed ACT scores would be represented by

$$ACT_1 = T_o + S_o + e_1 ,$$

and $ACT = T_o + S_o + e_2$ where e_1 and e_2 are uncorrelated with each other and with T_o and S_o. Further, T_o and S_o are uncorrelated. With the above model the reliability of the ACT composite might be closer to the expected value. This modification of the model, however, would not lead to changes in the other parameter estimates of the model shown in Figure 2.

The reliabilities shown in Figure 2 also may be lower for this sample of students who have completed eight semesters of college than it would be for a full range of high school students. The results are however consistent with the hypothesis that high school rank and ACT composite measure the same true variable and that the simplex model fits high school and college achievement.

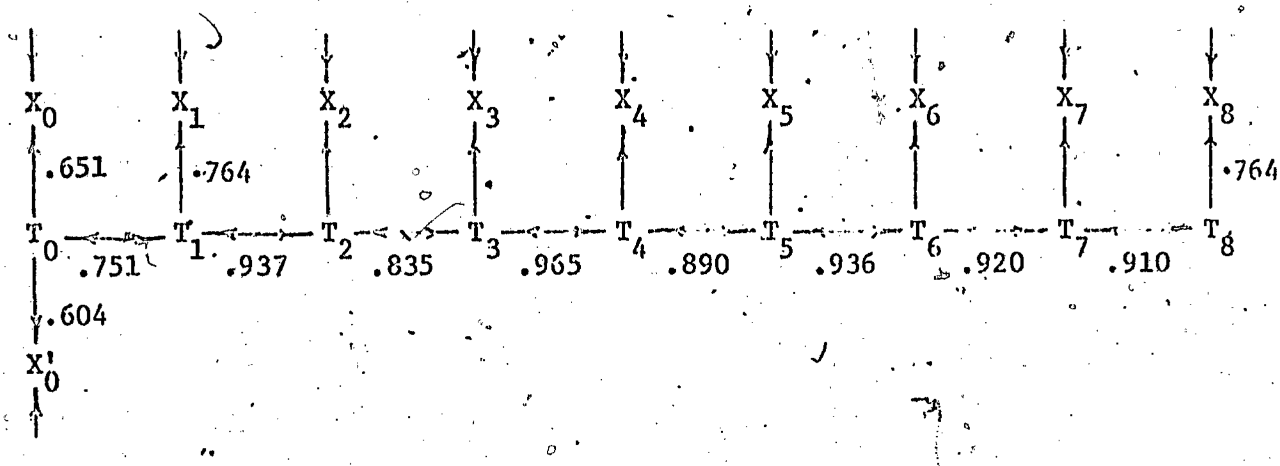


Figure 2. Structural model including high school achievement

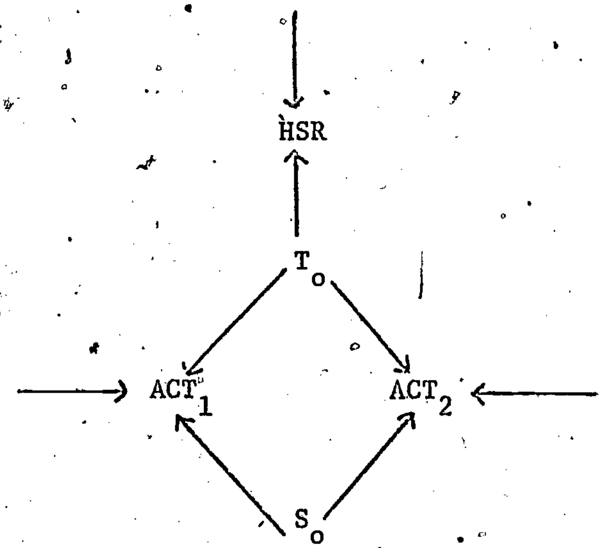


Figure 3. Postulated Model for HSR and Two ACT Composite Scores

D. Additional Analyses

Humphreys and Taber (1973, Table 1) provided correlations among eight semesters of college grades among seniors at the University of Illinois who took the Graduate Record Examination. Using a simplex model assuming equal reliabilities a chi square of 43.5 with 20 degrees of freedom was obtained. Although this is statistically significant at the P = .002 level the fit as judged from discrepancies between the observed correlations and those estimated from the model was close to that shown in Table 2. In part the poorer fit might have resulted from the fact that these are missing data correlations with sample sizes ranging from 1549 to 3018. The estimated reliability of .683 is somewhat higher than that previously estimated. It is interesting to note that the ratio of these reliabilities, i.e., .683 to .584, is approximately the ratio of the corresponding grade-point average variances, i.e., .380 to .331. Correlations between true scores from left to right in Figure 1 are .889, .939, .897, .942, and .909. These true correlations approximate those previously estimated.

III. Discussion

Jöreskog (1970a) found that a simplex model provided a fair fit to Humphreys' (1960) data on eight semesters grades in electrical engineering (N = 91). Combined with the present results it may be concluded that a simplex model provides a good fit to University of Illinois data. Whether this would be true for other institutions or for commingled grades from different schools remains to be demonstrated. The results support Humphreys' conclusion that the reliabilities across semesters were equal.

Bibliography

Humphreys, L. G. Investigations of the simplex. Psychometrika, 1960, 25, 313-323.

Humphreys, L. G. The fleeting nature of college academic success. Journal of Educational Psychology, 1968, 59, 375-380.

Humphreys, L. G., & Taber, T. Postdiction study of the graduate record examination and eight semesters of college grades. Journal of Educational Measurement, 1973, 10, 179-184.

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

Jöreskog, K. G. A general method for analysis of covariance structures. Biometrika, 1970, 57, 239-251. (b)

F. Incorporating Nonindependent Measurement Errors

This section resulted from efforts to deal with correlated errors of measurement. While it is written as if the problem were correlated ratings, the same problem arises when the same or similar tests are administered over time. The models in this section can be incorporated into the simplex framework by specifying the dimensions over time to have a simplex structure as shown in section C.

It is frequently the case that an expert is asked to rate the same objects along two or more dimensions. In these circumstances it is difficult for a judge to not let ratings on a dimension be influenced by knowledge of ratings on other dimensions. This kind of contamination means that the errors of measurement for one dimension may be correlated with the errors on other dimensions, i.e, the intrajudge measurement errors are correlated. Under these conditions, the covariance between ratings of different dimensions by the same judge is not equal to the covariance between the underlying true scores as would normally be assumed in classical test theory. The usual correction for attenuation formula for obtaining an estimate of the correlation between the underlying true scores on the dimensions would be inapplicable since uncorrelated errors are assumed in that formula. Presented herein is a procedure for analyzing data with correlated intrajudge and uncorrelated interjudge measurement errors. In addition to testing the fit of the model to the data, this procedure estimates correlations between the true scores on the dimensions, the reliabilities for each judge on each dimension, and the correlations between intrajudge errors.

I. Problem Formulation

Let S_{ij} be the rating of the i^{th} judge ($i = 1 \ldots N$) on the j^{th} dimension ($j = 1 \ldots M$).

Suppose that

$$X_{ij} = b_{ij}T_j + e_{ij} \quad (1)$$

where b_{ij} ≡ regression weight of X_{ij} on T_j ,

T_j ≡ true score for the $j^{\underline{th}}$ dimension,

and e_{ij} ≡ error of measurement for the $i^{\underline{th}}$ judge on the $j^{\underline{th}}$ dimension.

The e_{ij} are assumed to have a mean of zero and to be uncorrelated with the T_j. For convenience, the variance of T_j is set equal to unity. At this stage the model is similar to the traditional test theory model except that b_{ij} is not assumed to be the same for all judges as would be true in the case of parallel measures, and no assumption has been made about the correlation of the errors of measurement.

The allowance for correlated intrajudge measurement errors means that for a given i, errors (e_{ij}) for different values of j are correlated. It is assumed, however that the interjudge errors are uncorrelated which means that all e_{ij} for different values of i are uncorrelated.

A factor analytic model is appropriate for analyzing these data, however, because certain errors are correlated it is computationally convenient to treat the e_{ij} as factors along with the T_j. The dispersion matrix (Σ) of the X_{ij} has the form:

$$\Sigma = \Lambda \, \Phi \, \Lambda' \quad (2)$$

where Φ ≡ variance-covariance matrix among the factors (T_j and e_{ij}),

and Λ ≡ matrix of factor coefficients of the X_{ij} on the specified factors.

It is necessary to have at least three independent judges in order for the correlations among the T_j and error covariances to be uniquely estimated, i.e., for the model to be identified (Fisher, 1966). With only two judges the elements of Λ and Φ cannot be uniquely estimated and no test of the model fit is possible without additional assumptions.

Let S be the observed correlation matrix among X_{ij}. Fit of this model [i.e., equation (2)] will be judged by the deviation of the best fit estimate of Σ from S. For large samples it is also possible to test this fit.

II. The Three Judge, Two Dimension Model

Because of identification requirements it is expected that the three judge, two dimension model will be the basic building block for structures of this type.

The basic equations are:

$$X_{11} = b_{11}T_1 + e_{11} ,$$

$$X_{12} = b_{12}T_2 + e_{12} ,$$

$$X_{21} = b_{21}T_1 + e_{21} ,$$

$$X_{22} = b_{22}T_2 + e_{22} ,$$

$$X_{31} = b_{31}T_1 + e_{31} ,$$

and

$$X_{32} = b_{32}T_2 + e_{32} .$$

The vector of factors is $[T_1, T_2, e_{11}, e_{12}, e_{21}, e_{22}, e_{31}, e_{32}]$, and the Λ and Φ matrices in equation (2) have the following form:

$$\Lambda = \begin{bmatrix} b_{11} & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & b_{12} & 0 & 1 & 0 & 0 & 0 & 0 \\ b_{21} & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & b_{22} & 0 & 0 & 0 & 1 & 0 & 0 \\ b_{31} & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & b_{32} & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$

and

$$\Phi = \begin{bmatrix} 1 & & & & & & & \\ \Phi_0 & 1 & & & & \text{Symmetric} & & \\ 0 & 0 & \Phi_1 & & & & & \\ 0 & 0 & \Phi_2 & \Phi_3 & & & & \\ 0 & 0 & 0 & 0 & \Phi_4 & & & \\ 0 & 0 & 0 & 0 & \Phi_5 & \Phi_6 & & \\ 0 & 0 & 0 & 0 & 0 & 0 & \Phi_7 & \\ 0 & 0 & 0 & 0 & 0 & 0 & \Phi_8 & \Phi_9 \end{bmatrix}$$

where the b_{ij} and the Φ_r elements are parameters to be estimated.

The specifications for Φ indicate that the basic dimensions T_1 and T_2 are standardized by assigning a variance of unity to the corresponding diagonal elements, which means that the covariance of these factors $[\Phi_0]$

is a correlation. The six error variances [i.e., Φ_1, Φ_3, Φ_4, Φ_6, Φ_7 and Φ_9] and three intrajudge error covariances [i.e., Φ_2, Φ_5 and Φ_8] are to be estimated.

In order to explore the identifiability of parameters it is useful to perform the matrix multiplication indicated in equation (2) and examine the entries in Σ for the above specifications.

The diagonal entries of Σ are variances and are given by:

$$V(X_{ij}) = b_{ij}^2 + V(e_{ij}) \ . \tag{3}$$

The $V(e_{ij})$ are equal to particular diagonal elements in Φ ; for example, $V(e_{11}) = \Phi_1$, and $V(e_{21}) = \Phi_3$.

The off-diagonal elements of Σ that correspond to a single dimension (j = constant and $i \neq k$) are:

$$C(X_{ij}X_{kj}) = b_{ij}b_{kj} \ .$$

Given three judges it follows that

$$b_{1j}^2 = \frac{C(X_{1j}X_{2j})\ C(X_{1j}X_{3j})}{C(X_{2j}X_{3j})}, \tag{4}$$

$$b_{2j}^2 = \frac{C(X_{1j}X_{2j})\ C(X_{2j}X_{3j})}{C(X_{1j}X_{3j})},$$

and

$$b_{3j}^2 = \frac{C(X_{1j}X_{3j})\ C(X_{2j}X_{3j})}{C(X_{1j}X_{2j})} \ .$$

Since the b_{ij}^2 can be expressed in terms of the elements in Σ it follows that these factor coefficients (reliabilities) are identifiable. This in turn means that the $V(e_{ij})$ are identified from equation (3).

The off-diagonal elements corresponding to different judges and different dimensions ($i \neq k$, $j = 1, 2$) are:

$$C(X_{i1}X_{k2}) = b_{i1}b_{k2}C(T_1T_2) ,$$

where $C(T_1T_2) = \Phi_0$.

Since all b_{i1} and b_{k2} are identified as shown in (4), it follows that $C(T_1T_2)$ is identified.

Finally, the off-diagonal elements corresponding to a single judge (i = constant) and different dimensions are:

$$C(X_{i1}X_{i2}) = b_{i1}b_{i2}\ C(T_1T_2) + C(e_{i1}e_{i2})$$

where the $C(e_{i1}e_{i2})$ are equal to Φ_2, Φ_5 and Φ_8 for i equal 1, 2 and 3, respectively. Since b_{i1}, b_{i2}, and $C(T_1T_2)$ were shown above to be identifiable, it follows from these equations that the $C(e_{i1}e_{i2})$ are identified.

This model has 6 x 7 ÷ 2 = 21 unique elements in Σ and 16 model parameters (6 in Λ, 10 in Φ) which means that there are 21 - 16 = 5 degrees of overidentification. Overidentification is necessary to test the fit of the model to the data.

As specified above the error variance for a given rater and dimension is obtained from one of the diagonal entries of the Φ matrix. This formulation is convenient for investigating the question of identification as was done above. For purposes of estimation and interpretation, however,

$\lambda_1 = \sqrt{\phi_1}$,

$\lambda_2 = \sqrt{\phi_3}$,

$\lambda_3 = \sqrt{\phi_4}$,

$\lambda_4 = \sqrt{\phi_6}$,

$\lambda_5 = \sqrt{\phi_7}$,

and $\lambda_6 = \sqrt{\phi_9}$.

The ϕ^*'s are obtained from the ϕ's in the usual manner that a correlation is obtained from covariance and variance terms, e.g.

$$\phi_1^* = \frac{\phi_2}{\sqrt{\phi_1 \phi_3}}$$

It is the latter specification of Λ and ϕ that is used in the empirical example presented below.

III. Empirical Example

Jöreskog's (1970) general method for analysis of covariance structures and its associated computer program (Jöreskog, Gruvaeus, & van Thillo, 1970) were used for estimation. An earlier program for restricted maximum likelihood factor analysis (Jöreskog & Gruvaeus, 1967) or a more recent program for estimating a linear structural equation system (Jöreskog & van Thillo, 1972) could also be used. Details of the program are given in the manual.

To illustrate the computations, data provided by Dr. Donald Rock were used in which three judges rated thirty-four military positions on two dimensions [T_1 = Dealing with people and T_2 = Responsibility/Autonomy].

Intrajudge errors of measurement were probably correlated. The model for these data is that given in section II, above. The observed correlation matrix S is given in Table 1. The model was set up to yield correlations between the errors. Maximum likelihood estimates of model parameters in Λ and Φ are given in Table 2.

Table 1. The Observed Correlation Matrix

	X_{11}	X_{12}	X_{21}	X_{22}	X_{31}	X_{32}
X_{11}	1.0000					
X_{12}	.5851	1.0000				
X_{21}	.2462	.1218	1.0000			
X_{22}	.4110	.5360	.2709	1.0000		
X_{31}	.3823	.2946	.2033	.0694	1.0000	
X_{32}	.2816	.6114	.1675	.5049	.3314	1.0000

According to the model, reliability is defined as the square of the corresponding regression coefficient, e.g., the first dimension rater reliabilities were $(.766)^2$ = .587, .110, and .236, respectively. The correlation between underlying true dimensions is estimated as .605 and the correlations between intrajudge errors of measurement are .513, .177, and .265, respectively. The correlation (.513) between errors for the first judge approaches the true correlation (.605) between dimensions, indicating the necessity for methods which allow for such contingencies.

Table 2. Maximum Likelihood Parameter Estimates

$\hat{\Lambda}$ =

.766	0	.632	0	0	0	0	0
0	.792	0	.607	0	0	0	0
.331	0	0	0	.940	0	0	0
0	.671	0	0	0	.742	0	0
.486	0	0	0	0	0	.888	0
0	.769	0	0	0	0	0	.645

$\hat{\Phi}$ =

1.000						Symmetric	
.605	1.000						
0	0	1.000					
0	0	.513	1.000				
0	0	0	0	1.000			
0	0	0	0	.177	1.000		
0	0	0	0	0	0	1.000	
0	0	0	0	0	0	.265	1.000

A crucial part of the output is the estimated value of Σ and the corresponding discrepancies from the observed matrix S. If these are so large as to indicate that the data do not fit the model, then the above parameter estimates would have little meaning. The residials of $S - \Sigma$ are given in Table 3.

Table 3. Residuals $S-\Sigma$

	X_{11}	X_{12}	X_{21}	X_{22}	X_{31}	X_{32}
X_{11}	.013					
X_{12}	.021	.004				
X_{21}	-.007	-.037	.008			
X_{22}	.100	.005	.013	-.001		
X_{31}	.010	.062	.043	-.128	-.024	
X_{32}	-.075	.003	.014	-.011	-.046	-.007

Considering the relatively small sample size (only 34 persons rated by the three judges on the two dimensions) and the number of restrictions on the model, we interpret these results as a reasonably good fit of the model to the data. The goodness of fit test yielded a chi square of 4.54 with 5 degrees of freedom. The probability of getting a chi-squared value larger than that actually obtained, given that the hypothesized model is true, is P = .475. Since this chi-square test assumes a large sample, the small sample indicates that these results be interpreted with caution. In any event the chi-squared results do not indicate that the model should be rejected because of poor fit to the data.

V. Discussion

The model analyzed above was devised for the rating situation in which correlated measurement errors are likely. It may, however, provide an appropriate simulation in a variety of other situations, e.g.:

1. In the multitrait-multimethod procedure (Campbell & Fiske, 1959) the errors of measurement between two different trait measures using the same method may be correlated because of method variance. Method variance would be equivalent to correlated intrajudge errors when method factors are uncorrelated with trait factors.

2. In the use of the same test at two different times, the errors of measurement over time may be correlated because of practice and recall effects. At least three different measures of the underlying construct would be necessary for analysis. Such a model would be appropriate for the study of change over time by appropriate formulation of the factors (Werts, Jöreskog, & Linn, 1972).

The model does not explicitly state the causes of the correlation between the intrajudge errors, however it is assumed that the causes for each set of errors are uncorrelated with the causes for the other sets and of the dimensions being measured. A good fit of the model to the data implies that these assumptions are consistent with the results.

Insight may be gained into the meaning of a good fit with the data by examining the equations of the three judge, two dimension model for the variables X_{11}, X_{21}, X_{31} and X_{12}. In section II it was demonstrated that the three measures of T_1, i.e., X_{11}, X_{21}, and X_{31} identify the three regression weights of ratings on T_1:

$$b_{11}^2 = \frac{C(X_{11}X_{21})\ C(X_{11}X_{31})}{C(X_{21}X_{31})}$$

$$b_{21}^2 = \frac{C(X_{11}X_{21})\ C(X_{21}X_{31})}{C(X_{11}X_{31})}, \text{ and}$$

$$b_{31}^2 = \frac{C(X_{11}X_{31})\ C(X_{21}X_{31})}{C(X_{11}X_{21})}.$$

In factor analytic language, given three measures of a single factor with uncorrelated residuals, the three factor loadings may be uniquely estimated. However with only three measures of a factor, no test of the assumption of single factoredness is possible because a perfect fit with the data is always achieved (although communalities greater than one may be required). Thus, the model has no degrees of overidentification. It is of interest therefore to examine the relationship among X_{21}, X_{31} and X_{12} which yield:

$$b_{21}^{\;2} = \frac{C(X_{12}X_{21})\ C(X_{21}X_{31})}{C(X_{12}X_{31})}$$

and

$$b_{31}^{\;2} = \frac{C(X_{12}X_{31})\ C(X_{21}X_{31})}{C(X_{12}X_{21})} .$$

Comparison of the pairs of equations for $b_{21}^{\;2}$ and $b_{31}^{\;2}$ indicates that for the purposes of identifying and estimating these loadings, X_{12} is functionally equivalent to X_{11}. In other words, even though X_{12} is not a measure of T_1, it nevertheless allows a test of the hypothesis of single-factoredness of T_1. The basic reason for this is that e_{12} is uncorrelated with T_1, e_{21}, and e_{31} even though it is correlated with e_{11}. The finding of a good fit to the data is therefore consistent with the assumption that the observed variables are in fact measures of the specified dimension. A poor fit might be due to the falsity of this assumption, however one or more of the other model assumptions may be erroneous.

The variety of models involving correlated errors is too great to be detailed herein. For most of these the three judge, two dimension model is likely to be the basic unit. Within the constraints set by the computer program, the available data should, however, be analyzed by a single model. For example, three judge, three dimension data could be computed using the three judge, two dimension model for each of the three different pairs of dimensions. The result would be that for each reliability two estimates would be obtained which might differ considerably. A simultaneously estimated three judge, three dimension model would yield a single best fit estimate for all the data. Providing a good fit is obtained, a parameter derived from the three judge, three dimension model should have greater generalizability because it implies that the two estimates from the corresponding three judge, two dimension models are consistent.

Campbell, D. T., & Fiske, D. W. Convergent and discriminant validation by validation by the multitrait-multimethod matrix. Psychological Bulletin, 1959, 56, 81-105.

Fisher, F. M. The identification problem in econometrics. New York: McGraw-Hill, 1966.

Jöreskog, K. G., & Gruvaeus, G. T. RMLFA - A computer program for restricted maximum likelihood factor analysis. RB-67-21. Princeton, N. J.: Educational Testing Service, 1967.

Jöreskog, K. G. A general method for analysis of covariance structures. Biometrika, 1970, 57, 239-251.

Jöreskog, K. G., Gruvaeus, G. T., & van Thillo, M. ACOVS - A general computer program for analysis of covariance structures. RB-70-15. Princeton, N. J.: Educational Testing Service, 1970.

Jöreskog, K. G., & van Thillo, M. LISREL - A general computer program for estimating a linear structural equation system involving multiple indicators of unmeasured variables. RB-72-56. Princeton, N. J.: Educational Testing Service, 1972.

Werts, C. E., Joreskog, K. G., & Linn, R. L. A multitrait-multimethod model for studying growth. Educational and Psychological Measurement, 1972, 32, 655-678.

G. Conclusions

The most important finding of this study is that a simplex model which allows for measurement error, fits a variety of longitudinal academic data quite well. As detailed in section D, this allows for attenuation corrections when only one measure is available at each time. More importantly, the results suggest that the commonly used split-half or parallel form procedures for estimating reliability may typically yield overestimates of reliability due to "method" variance i.e., nonindependent measurement errors resulting from the use of closely similar item types. The simplex model appears less subject to this problem because both item formats and content change over time. It has been demonstrated that accurate corrections for attenuation are essential to a study of the determinants of academic growth.

These initial results have encouraged us to incorporate the simplex model into larger structural models, with favorable results. Furthermore, the problem of combining simplexes for different measures was found feasible. These results will be forthcoming in the literature as soon as the analyses are completed.