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)
STUDY OF ACADEMIC GROWTH USING SIMPLEX MODELS

Charles E. Weerts 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.
## Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Objectives</td>
<td>3</td>
</tr>
<tr>
<td>B. Introduction</td>
<td>4</td>
</tr>
<tr>
<td>C. Methodology of a Simplex Growth Model</td>
<td>5</td>
</tr>
<tr>
<td>D. Implications for the Study of Growth</td>
<td>22</td>
</tr>
<tr>
<td>E. Analyses of Longitudinal Grade Data</td>
<td>35</td>
</tr>
<tr>
<td>F. Incorporating Nonindependent Measurement Errors</td>
<td>48</td>
</tr>
<tr>
<td>G. Conclusion</td>
<td>64</td>
</tr>
</tbody>
</table>
Acknowledgments

The authors are indebted to Dr. Karl Jureškoň, 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; Lunnéborg & Lunnéborg, 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 \( \eta_i \) by the traditional equation:

\[
y_i = \eta_i + \epsilon_i
\]

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

\[
\eta_{i+1} = B_1 \eta_i + \epsilon_{i+1}
\]

where all \( \epsilon_i \) are independent and \( B_1 \) is the true regression weight. As noted by Humphreys (1960), equation (2) implies that the partial correlation between \( \eta_i \) and \( \eta_{i+2} \) is zero with \( \eta_{i+1} \) controlled. The special characteristic of a growth model (Humphreys, 1960) is that successive \( \eta_i \) have the same units of measurement and the difference \( \Delta_i \) between successive \( \eta_i \) is given by:

\[
\eta_{i+1} = \eta_i + \Delta_i
\]

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

\[
\Delta_i = (B_1 - 1)\eta_i + \epsilon_{i+1}
\]

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

\[
B\eta = \Gamma\xi + \zeta
\]

(5)

where \( B(m \times m) \) and \( \Gamma(m \times n) \) are coefficient matrices and \( \xi' = (\xi_1', \xi_2', \ldots, \xi_n) \) is a random vector of residuals (errors in equations, random disturbance terms). Without loss of generality it may be assumed that \( \mathcal{E}(\eta) = \mathcal{E}(\xi) = 0 \) and \( \mathcal{E}(\zeta) = 0 \). It is furthermore assumed that \( \zeta \) is uncorrelated with \( \xi \) and that \( B \) is nonsingular.

The vectors \( \eta \) and \( \xi \) are not observed but instead vectors \( y' = (y_1, y_2, \ldots, y_p) \) and \( x' = (x_1, x_2, \ldots, x_q) \) are observed, such that

\[
y = y' + \Lambda_y \eta + \epsilon
\]

(6)

\[
x = x' + \Lambda_x \xi + \delta
\]

(7)

where \( y = \mathcal{E}(y) \), \( x = \mathcal{E}(x) \) and \( \epsilon \) and \( \delta \) 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 \( \eta \) and of \( x \) on \( \xi \), respectively. It is convenient to refer to \( y \) and \( x \) as the observed variables and \( \eta \) and \( \xi \) as the true variables. The errors of measurement are assumed to be uncorrelated with each other and with the true variables.

Let \( \Phi(m \times n) \) and \( \Psi(m \times m) \) be the variance-covariance matrices of \( \xi \) and \( \xi' \), respectively, \( \Theta_\epsilon \) and \( \Theta_\delta \) 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 \( \zeta = (y', x')' \) is

\[
\Sigma = \begin{pmatrix}
\Lambda_y(B^{-1}\Phi'B^{-1} + B^{-1}\Psi'B^{-1})\Lambda_y + \Theta_\epsilon & \Lambda_yB^{-1}\Phi A' - \Lambda_yB^{-1}\Phi \Lambda_x + \Theta_\epsilon \\
\Lambda_xB^{-1}\Phi A' + \Theta_\delta & \Lambda_x + \Theta_\delta
\end{pmatrix}
\]

(8)

The elements of \( \Sigma \) are functions of the elements of \( \Lambda_y \), \( \Lambda_x \), \( B \), \( \Gamma \), \( \Phi \), \( \psi \), \( \Theta_\epsilon \), and \( \Theta_\delta \). In applications some of these elements are fixed and equal to assigned values. In particular this is so for elements in \( \Lambda_y \), \( \Lambda_x \), \( B \) and \( \Gamma \), 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 \( \Lambda_y \), \( \Lambda_x \), \( B \), \( \Gamma \), \( \Phi \), \( \psi \), \( \Theta_\epsilon \), and \( \Theta_\delta \) are of three kinds: (i) fixed parameters that have been assigned given values, (ii) constrained parameters that are unknown but equal to one or more other parameters and (iii) free parameters 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 $A_x$, $\Gamma$, $G$, and $D_0$ are not required and may be deleted using a program option which specifies no $x$. Comparison of equations (2) and (6) indicates that $A_y$ is an identity matrix. Equation (2) must be changed to $-B_1 \eta_1 + \eta_{i+1} = \zeta_{i+1}$ to be in equation (5) format with $\Gamma$ and $G$ 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

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

e. In equation (6) $A_y$ is an 8 by 8 identity matrix.
f. In equation (5)

\[
B = \begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-B_1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & -B_2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & -B_3 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & -B_4 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & -B_5 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & -B_6 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & -B_7 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & -B_8 & 1 \\
\end{bmatrix}
\]

g. The variance covariance matrix, \( \Psi \), of the \( \xi_i \) is a diagonal matrix with entries \( V_{\xi_i} \) where \( i = 1, 2, \ldots, 8 \).

h. The variances of the \( \xi_i \) are

\[
\Theta^2 = [V_{\xi_1}, V_{\xi_2}, V_{\xi_3}, V_{\xi_4}, V_{\xi_5}, V_{\xi_6}, V_{\xi_7}, V_{\xi_8}].
\]

Following Jöreskog (1970) it can be shown that \( V_{\xi_1}, V_{\xi_8}, B_1, V_{\xi_1}, V_{\xi_2} \) and \( V_{\xi_8} \) are not identified; i.e., unique estimates cannot be obtained. Identification of all parameters was achieved by arbitrarily assigning fixed values to \( V_{\xi_1} \) and \( V_{\xi_8} \). Given these additional specifications, there are \( 8 \times 9 / 2 = 36 \) distinct elements in \( \Theta \) and 21 parameters to be estimated (6 \( V_{\xi_1}, 7 B_1, \) and 8 \( V_{\xi_1} \)), which leaves 15 degrees of freedom (overidentifying restrictions) to test the fit of the model to the data.
The observed variance-covariance matrix \( \Sigma \) for the eight variables in the Bracht and Hopkins (1972) data is given in Table 1.

Table 1. Observed variance-covariance matrix

<table>
<thead>
<tr>
<th>( y_1 )</th>
<th>( y_2 )</th>
<th>( y_3 )</th>
<th>( y_4 )</th>
<th>( y_5 )</th>
<th>( y_6 )</th>
<th>( y_7 )</th>
<th>( y_8 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.260</td>
<td>0.257</td>
<td>0.336</td>
<td>0.371</td>
<td>0.416</td>
<td>0.437</td>
<td>0.465</td>
<td>1.576</td>
</tr>
<tr>
<td>0.257</td>
<td>0.476</td>
<td>0.528</td>
<td>0.451</td>
<td>0.646</td>
<td>0.661</td>
<td>0.705</td>
<td>2.168</td>
</tr>
<tr>
<td>0.336</td>
<td>0.528</td>
<td>0.792</td>
<td>0.782</td>
<td>0.918</td>
<td>0.942</td>
<td>0.995</td>
<td>3.208</td>
</tr>
<tr>
<td>0.371</td>
<td>0.551</td>
<td>0.782</td>
<td>1.020</td>
<td>1.127</td>
<td>1.158</td>
<td>1.213</td>
<td>4.005</td>
</tr>
<tr>
<td>0.416</td>
<td>0.646</td>
<td>0.918</td>
<td>1.127</td>
<td>1.440</td>
<td>1.158</td>
<td>1.490</td>
<td>5.006</td>
</tr>
<tr>
<td>0.437</td>
<td>0.661</td>
<td>0.942</td>
<td>1.158</td>
<td>1.406</td>
<td>1.406</td>
<td>1.634</td>
<td>5.516</td>
</tr>
<tr>
<td>0.465</td>
<td>0.705</td>
<td>0.995</td>
<td>1.213</td>
<td>1.490</td>
<td>1.490</td>
<td>1.904</td>
<td>6.112</td>
</tr>
</tbody>
</table>

LISREL yielded the following maximum-likelihood parameter estimates:

- Parameters in \( B : B_2 = 1.318, B_3 = 1.054, B_4 = 1.171, B_5 = 1.026, B_6 = 1.055 \) and \( B_7 = 3.367 \).
- Parameters in \( \psi : \epsilon_3 = .049, \epsilon_4 = .137, \epsilon_5 = .052, \epsilon_6 = .107, \epsilon_7 = .091 \).
- Parameters in \( \Theta : \epsilon_2 = (.276)^2, \epsilon_3 = (.222)^2, \epsilon_4 = (.240)^2, \epsilon_5 = (.260)^2, \epsilon_6 = (.193)^2, \epsilon_7 = (.298)^2 \).

Parameters not listed are not identified. The program also estimates the variance-covariance matrix among the true variables \( \{y_i\} \).
A crucial part of the output is the estimated value of $\xi$ 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 $S - \hat{S}$ are given in Table 2. Because of the difficulty in comparing residuals between variables with different variances, Table 2 gives discrepancies after $S$ and $\hat{S}$ have been standardized to correlation matrices. Considering the fairly large number of over-identifying 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 $S - \hat{S}$, standardized

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.000</td>
<td>0.000</td>
<td>-0.006</td>
<td>0.002</td>
<td>-0.000</td>
</tr>
<tr>
<td>0.027</td>
<td>-0.007</td>
<td>-0.001</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>-0.003</td>
<td>-0.005</td>
<td>0.001</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>0.013</td>
<td>-0.007</td>
<td>0.001</td>
<td>0.001</td>
<td>-0.001</td>
</tr>
<tr>
<td>0.018</td>
<td>0.002</td>
<td>0.003</td>
<td>-0.005</td>
<td>0.004</td>
</tr>
<tr>
<td>0.021</td>
<td>-0.056</td>
<td>-0.028</td>
<td>-0.019</td>
<td>0.001</td>
</tr>
</tbody>
</table>
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 $\eta_2$ and between $y_6$ and $\eta_7$ 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 Brächt & 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 Brächt & 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.

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

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

b. The true correlation of status with gain,
\(\rho(\eta_i, \Delta_1)\) is given by:
\(\hat{\rho}(\eta_i, \Delta_1) = (\hat{\beta}_1)^{-1} \sqrt{\hat{V}_{\eta_i} \cdot \hat{V}_{\Delta 1}}\) (10), and

c. the reliability of gain scores, \(\rho_{\Delta 1}\), is given by:
\(\hat{\rho}_{\Delta 1} = \frac{\hat{V}_{\Delta 1}}{\sqrt{\hat{V}_{\Delta 1} + \hat{V}_{\epsilon_1} + \hat{V}_{\epsilon_{i+1}}}\). (11).

The estimated variance-covariance matrix among the true variables is given in table 3 except for the unidentified variances of \(\eta_1\) and \(\eta_8\). For this reason no growth statistics involving \(\eta_1\) and \(\eta_8\) 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

<table>
<thead>
<tr>
<th>n1</th>
<th>n2</th>
<th>n3</th>
<th>n4</th>
<th>n5</th>
<th>n6</th>
<th>n7</th>
<th>n8</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.257</td>
<td>0.400</td>
<td>0.743</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.338</td>
<td>0.527</td>
<td>0.783</td>
<td>0.962</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.357</td>
<td>0.555</td>
<td>0.917</td>
<td>1.127</td>
<td>1.372</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.418</td>
<td>0.650</td>
<td>0.940</td>
<td>1.156</td>
<td>1.407</td>
<td>1.550</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.428</td>
<td>0.703</td>
<td>0.992</td>
<td>1.219</td>
<td>1.484</td>
<td>1.635</td>
<td>1.815</td>
<td></td>
</tr>
</tbody>
</table>

*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 (σ²). Table 3 could be used with equations (9), (10), and (11) to compute change statistics between any two true variables e.g. the n7 - n2 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

<table>
<thead>
<tr>
<th>Change</th>
<th>Change Variance</th>
<th>$b_1 - 1$</th>
<th>Status - Gain Correlation</th>
<th>Reliability of Gain</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\eta_3 - \eta_2$</td>
<td>.089</td>
<td>.318</td>
<td>.674</td>
<td>.415</td>
</tr>
<tr>
<td>$\eta_4 - \eta_3$</td>
<td>.139</td>
<td>.054</td>
<td>.125</td>
<td>.565</td>
</tr>
<tr>
<td>$\eta_5 - \eta_4$</td>
<td>.081</td>
<td>.171</td>
<td>.589</td>
<td>.393</td>
</tr>
<tr>
<td>$\eta_6 - \eta_5$</td>
<td>.108</td>
<td>.026</td>
<td>.093</td>
<td>.507</td>
</tr>
<tr>
<td>$\eta_7 - \eta_6$</td>
<td>.096</td>
<td>.055</td>
<td>.221</td>
<td>.432</td>
</tr>
</tbody>
</table>
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
\[ \eta_{t+1} = \eta_t + A_i \] in which case the vector of true variables becomes:
\[ \eta = (\eta_1, \Delta_1, \Delta_2, \Delta_3, \Delta_4, \Delta_5, \Delta_6, \Delta_7) \] and
\[ y, \varepsilon, \text{ and } \xi \] remain the same as before. Equation (1) becomes:
\[ y_i = (\eta_1 + \sum \Delta_i) \varepsilon_i \] (12)
and equation (4):
\[ \Delta_i = (B_i - 1)(\eta_1 + \sum \Delta_i) \varepsilon_i + \xi_{i+1} \] (13)

Translating these equations into equations (5) and (6):
\[ \begin{align*}
\Delta_i &= (B_i - 1)(\eta_1 + \sum \Delta_i) \\
&= (B_i - 1)\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
\end{bmatrix} \end{align*} \]

Translating equation (13) and defining \( b_i = (B_i - 1) \):
\[ b_i = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ b_1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ b_2 & b_1 & 1 & 0 & 0 & 0 & 0 & 0 \\ b_3 & b_2 & b_1 & 1 & 0 & 0 & 0 & 0 \\ b_4 & b_3 & b_2 & b_1 & 1 & 0 & 0 & 0 \\ b_5 & b_4 & b_3 & b_2 & b_1 & 1 & 0 & 0 \\ b_6 & b_5 & b_4 & b_3 & b_2 & b_1 & 1 & 0 \\ b_7 & b_6 & b_5 & b_4 & b_3 & b_2 & b_1 & 1 \\
\end{bmatrix} \]
This model is simply a linear transformation of the previous model, has the same number of parameters to be estimated \((V_1 \& V_2, \text{arbitrarily fixed})\), 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 \(B_3\)) of the LISREL option permitting parameters to be specified as equal.

The variance-covariance matrix of the true variables (\(C\)) in the growth formulation yields the estimated true change variance \((V_{\Delta}^m)\) directly.

It is reasonable to use this formulation in conjunction with the previous formulation to obtain the change statistics. Also of supplemental interest is \(C_{\text{standardized}}\) which gives the correlation among the change factors as shown in Table 5 (\(A_1\) and \(A_7\) 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

<table>
<thead>
<tr>
<th>(A_2)</th>
<th>(A_3)</th>
<th>(A_4)</th>
<th>(A_5)</th>
<th>(A_6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.104</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.460</td>
<td>.290</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>.069</td>
<td>.044</td>
<td>.067</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>.162</td>
<td>.102</td>
<td>.156</td>
<td>.077</td>
<td>1.000</td>
</tr>
</tbody>
</table>
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 \( \Delta_4 \) are given by:

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

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

\[
C(y_3, \Delta_4) = [B_4 - 1] B_3 \eta_3
\]

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

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

\[
C(y_6, \Delta_4) = B_5 C(y_5, \Delta_4)
\]

\[
C(y_7, \Delta_4) = B_6 C(y_6, \Delta_4)
\]

\[
C(y_8, \Delta_4) = B_7 C(y_7, \Delta_4)
\]

For the purpose of estimating \( \Delta_4 \) it is appropriate to use the estimated variances and covariances among the \( y_i \) as obtained from \( \Sigma \) rather than the observed matrix because \( \Sigma \) 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 \( \Delta_4 \) on the \( y_i \). The \( y_i \) predicted 63% of the variance in \( \Delta_4 \). This is not much larger than the 59% of the
Table 6. Correlations Among $y_1$ and $\Delta_4$

<table>
<thead>
<tr>
<th>$\Delta_4$</th>
<th>$y_1$</th>
<th>$y_2$</th>
<th>$y_3$</th>
<th>$y_4$</th>
<th>$y_5$</th>
<th>$y_6$</th>
<th>$y_7$</th>
<th>$y_8$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y_1$</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$y_2$</td>
<td>.421</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$y_3$</td>
<td>.484</td>
<td>.731</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$y_4$</td>
<td>.529</td>
<td>.747</td>
<td>.858</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$y_5$</td>
<td>.572</td>
<td>.693</td>
<td>.797</td>
<td>.871</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$y_6$</td>
<td>.719</td>
<td>.683</td>
<td>.785</td>
<td>.859</td>
<td>.970</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$y_7$</td>
<td>.702</td>
<td>.666</td>
<td>.767</td>
<td>.838</td>
<td>.908</td>
<td>.930</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>$y_8$</td>
<td>.677</td>
<td>.642</td>
<td>.738</td>
<td>.808</td>
<td>.875</td>
<td>.896</td>
<td>.940</td>
<td>1.000</td>
</tr>
</tbody>
</table>
variance predictable from \( y_4 \) and \( y_5 \) alone because the reliabilities of \( y_4 \) and \( y_5 \) are quite high. Note that the \( \Delta \) 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 \( \eta_5 \). 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 \((m \times n = 12)\), 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


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 \( E_i \):

\[ X_i = T_i + E_i \]  \hspace{1cm} (1)

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

\[ T_{i+1} = B_{i} T_i + d_i, \]  \hspace{1cm} (2)

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

\[ R_{ii} = \frac{R_{in} R_{im}}{R_{nm}} \]  \hspace{1cm} (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 ($\Delta_i$) 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 ($\beta_i$), 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 education 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.
<table>
<thead>
<tr>
<th>TEST</th>
<th>$V_1$</th>
<th>$V_2$</th>
<th>$V_3$</th>
<th>$V_4$</th>
<th>$C_{12}$</th>
<th>$C_{13}$</th>
<th>$C_{14}$</th>
<th>$C_{23}$</th>
<th>$C_{24}$</th>
<th>$C_{34}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Science</td>
<td>150.587</td>
<td>121.895</td>
<td>185.270</td>
<td>160.501</td>
<td>97.724</td>
<td>116.270</td>
<td>98.472</td>
<td>110.890</td>
<td>90.232</td>
<td>121.433</td>
</tr>
<tr>
<td>Social S.</td>
<td>130.444</td>
<td>172.867</td>
<td>222.648</td>
<td>216.681</td>
<td>117.774</td>
<td>126.673</td>
<td>117.584</td>
<td>154.653</td>
<td>146.644</td>
<td>168.819</td>
</tr>
<tr>
<td>Reading</td>
<td>252.143</td>
<td>299.958</td>
<td>256.291</td>
<td>298.194</td>
<td>210.990</td>
<td>183.106</td>
<td>195.151</td>
<td>216.601</td>
<td>222.691</td>
<td>212.424</td>
</tr>
<tr>
<td>Listening</td>
<td>150.438</td>
<td>195.429</td>
<td>214.374</td>
<td>216.652</td>
<td>132.970</td>
<td>130.916</td>
<td>120.145</td>
<td>161.556</td>
<td>146.774</td>
<td>156.870</td>
</tr>
<tr>
<td>Writing</td>
<td>198.500</td>
<td>261.488</td>
<td>313.265</td>
<td>320.804</td>
<td>173.126</td>
<td>180.092</td>
<td>175.054</td>
<td>222.842</td>
<td>213.632</td>
<td>239.027</td>
</tr>
<tr>
<td>SCAT:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Verbal</td>
<td>135.320</td>
<td>157.242</td>
<td>183.920</td>
<td>208.363</td>
<td>124.718</td>
<td>127.706</td>
<td>132.553</td>
<td>149.294</td>
<td>153.476</td>
<td>171.799</td>
</tr>
<tr>
<td>Quant.</td>
<td>73.000</td>
<td>164.858</td>
<td>265.328</td>
<td>310.609</td>
<td>80.160</td>
<td>96.224</td>
<td>100.512</td>
<td>157.130</td>
<td>165.122</td>
<td>231.356</td>
</tr>
<tr>
<td>Total</td>
<td>58.722</td>
<td>104.044</td>
<td>157.417</td>
<td>199.128</td>
<td>66.520</td>
<td>78.259</td>
<td>85.104</td>
<td>112.751</td>
<td>122.638</td>
<td>159.112</td>
</tr>
</tbody>
</table>

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 parameter 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 variance-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.
<table>
<thead>
<tr>
<th>TEST</th>
<th>BASIC $\chi^2$</th>
<th>BASIC P</th>
<th>RMS Residuals</th>
</tr>
</thead>
<tbody>
<tr>
<td>STEP:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Math</td>
<td>0.17</td>
<td>.683</td>
<td>.001</td>
</tr>
<tr>
<td>Science</td>
<td>5.39*</td>
<td>.020</td>
<td>.005</td>
</tr>
<tr>
<td>Social S.</td>
<td>3.09</td>
<td>.079</td>
<td>.003</td>
</tr>
<tr>
<td>Reading</td>
<td>8.06*</td>
<td>.005</td>
<td>.004</td>
</tr>
<tr>
<td>Listening</td>
<td>0.53</td>
<td>.468</td>
<td>.001</td>
</tr>
<tr>
<td>Writing</td>
<td>1.06</td>
<td>.303</td>
<td>.002</td>
</tr>
<tr>
<td>SCAT:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Verbal</td>
<td>2.18</td>
<td>.140</td>
<td>.001</td>
</tr>
<tr>
<td>Quant.</td>
<td>0.24</td>
<td>.623</td>
<td>.001</td>
</tr>
<tr>
<td>Total</td>
<td>0.00</td>
<td>.973</td>
<td>.000</td>
</tr>
</tbody>
</table>

Table 2. Chi-square Goodness-of-Fit Tests

*Significant at the .02 level or better when the $\chi^2$ 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 \( V_{T1}, V_{T4}, B_1, R_{11}, R_{44}, V_{A1}, V_3, \rho_{T1A1}, \rho_{T3A3} \) 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 \( \rho_{T2A2} \) in the last column of Table 3. The true change variance from seventh to ninth grades \( V_{A2} \) can be compared to the sum of the error variances for the tests at these times (labelled \( V_{AE} \) in Table 3).
The reliability of these gain scores would be $V_{\Delta r} = V_{\Delta 2} + V_{\Delta E}$. 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.
<table>
<thead>
<tr>
<th>TEST</th>
<th>$VT_2$</th>
<th>$VT_3$</th>
<th>$B_2$</th>
<th>$B_3$</th>
<th>$R_{22}$</th>
<th>$R_{33}$</th>
<th>$V_{\Delta 2}$</th>
<th>$V_{\Delta E}$</th>
<th>$p_{T2\Delta 2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Math</td>
<td>135.000</td>
<td>147.771</td>
<td>1.010</td>
<td>1.047*</td>
<td>.753</td>
<td>.802</td>
<td>10.097</td>
<td>81.032</td>
<td>+.037</td>
</tr>
<tr>
<td>Science</td>
<td>92.069</td>
<td>146.752</td>
<td>1.198*</td>
<td>0.827*</td>
<td>.755</td>
<td>.792</td>
<td>18.167</td>
<td>68.167</td>
<td>+.446</td>
</tr>
<tr>
<td>Social S.</td>
<td>145.015</td>
<td>179.352</td>
<td>1.070*</td>
<td>0.941*</td>
<td>.839</td>
<td>.806</td>
<td>14.151</td>
<td>71.154</td>
<td>+.224</td>
</tr>
<tr>
<td>Reading</td>
<td>245.792</td>
<td>203.956</td>
<td>0.876*</td>
<td>1.042*</td>
<td>.819</td>
<td>.796</td>
<td>18.912</td>
<td>106.515</td>
<td>-.447</td>
</tr>
<tr>
<td>Listening</td>
<td>163.556</td>
<td>172.089</td>
<td>0.987</td>
<td>0.912*</td>
<td>.837</td>
<td>.803</td>
<td>12.883</td>
<td>74.417</td>
<td>-.046</td>
</tr>
<tr>
<td>Writing</td>
<td>213.038</td>
<td>248.105</td>
<td>1.044*</td>
<td>0.963*</td>
<td>.815</td>
<td>.792</td>
<td>16.337</td>
<td>113.613</td>
<td>+.159</td>
</tr>
<tr>
<td>SCAT:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Verbal</td>
<td>145.285</td>
<td>166.737</td>
<td>1.027*</td>
<td>1.030*</td>
<td>.924</td>
<td>.907</td>
<td>13.684</td>
<td>29.139</td>
<td>+.088</td>
</tr>
<tr>
<td>Quant.</td>
<td>132.069</td>
<td>221.339</td>
<td>1.197*</td>
<td>1.049*</td>
<td>.796</td>
<td>.831</td>
<td>37.352</td>
<td>78.238</td>
<td>+.370</td>
</tr>
<tr>
<td>Total</td>
<td>95.846</td>
<td>146.293</td>
<td>1.176*</td>
<td>1.088*</td>
<td>.921</td>
<td>.929</td>
<td>16.631</td>
<td>19.319</td>
<td>+.423</td>
</tr>
</tbody>
</table>

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, .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.


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 \( \epsilon_i \) of measurement:

\[
X_i = T_i + \epsilon_i \quad (1)
\]

All of the \( \epsilon_i \) are assumed independent and successive \( T_i \) are related by the linear regression equation:

\[
T_{i+1} = \beta_i T_i + d_i \quad (2)
\]

where the \( d_i \) residuals are assumed independent of each other.

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

\[
r_{ii} = \frac{r_{in} r_{im}}{r_{nm}} \quad (3)
\]
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

<table>
<thead>
<tr>
<th></th>
<th>$X_0$</th>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$X_4$</th>
<th>$X_5$</th>
<th>$X_6$</th>
<th>$X_7$</th>
<th>$X_8$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_0$</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$X_1$</td>
<td>0.393</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$X_2$</td>
<td>0.387</td>
<td>0.375</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$X_3$</td>
<td>0.341</td>
<td>0.298</td>
<td>0.556</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$X_4$</td>
<td>0.278</td>
<td>0.237</td>
<td>0.456</td>
<td>0.490</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$X_5$</td>
<td>0.270</td>
<td>0.255</td>
<td>0.439</td>
<td>0.445</td>
<td>0.562</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$X_6$</td>
<td>0.240</td>
<td>0.238</td>
<td>0.415</td>
<td>0.418</td>
<td>0.496</td>
<td>0.512</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$X_7$</td>
<td>0.256</td>
<td>0.252</td>
<td>0.399</td>
<td>0.383</td>
<td>0.456</td>
<td>0.469</td>
<td>0.551</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>$X_8$</td>
<td>0.240</td>
<td>0.219</td>
<td>0.387</td>
<td>0.364</td>
<td>0.445</td>
<td>0.442</td>
<td>0.500</td>
<td>0.544</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Note: $X_0$ is high school rank, $X_1\,\text{ACT}$ 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 $X_1$ and $T_1$ 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. SimpIex parameter estimates
Table 2
Discrepancies between Observed and Estimated Correlations among Observed Variables

<table>
<thead>
<tr>
<th></th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>X5</th>
<th>X6</th>
<th>X7</th>
<th>X8</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1</td>
<td>-0.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X2</td>
<td>-0.002</td>
<td>0.004</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X3</td>
<td>-0.011</td>
<td>0.008</td>
<td>0.002</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X4</td>
<td>-0.009</td>
<td>-0.018</td>
<td>0.008</td>
<td>0.008</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X5</td>
<td>0.012</td>
<td>0.002</td>
<td>-0.002</td>
<td>-0.002</td>
<td>-0.002</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X6</td>
<td>0.023</td>
<td>-0.006</td>
<td>-0.010</td>
<td>-0.011</td>
<td>0.007</td>
<td>-0.001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>X7</td>
<td>0.037</td>
<td>0.003</td>
<td>0.012</td>
<td>-0.004</td>
<td>-0.005</td>
<td>0.005</td>
<td>-0.010</td>
<td></td>
</tr>
<tr>
<td>X8</td>
<td>0.025</td>
<td>0.012</td>
<td>-0.038</td>
<td>0.012</td>
<td>-0.005</td>
<td>0.006</td>
<td>0.007</td>
<td>-0.000</td>
</tr>
</tbody>
</table>
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 $X_1$ and $T_1$ 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.,

\[
X_0 = T_0 + e_0 \quad \text{and} \quad X'_0 = T_0 + e'_0
\]

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 \( e_0 \) and \( e'_0 \)) 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 of 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 $e$ 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

$$\text{ACT}_1 = T_0 + S_0 + e_1$$

and $\text{ACT} = T_0 + S_0 + e_2$ where $e_1$ and $e_2$ are uncorrelated with each other and with $T_0$ and $S_0$. Further, $T_0$ and $S_0$ 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 1749 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 .900. 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


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

\[ x_{ij} = b_{ij} T_j + e_{ij} \]  

(1)

where \( b_{ij} \) = regression weight of \( x_{ij} \) on \( T_j \),

\( T_j \) = true score for the \( j^{th} \) dimension,

and \( e_{ij} \) = error of measurement for the \( i^{th} \) judge on the \( j^{th} \) 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 \( (\Sigma) \) of the \( x_{ij} \) has the form:

\[ \Sigma = \Lambda \phi \Lambda' \]  

(2)

where \( \phi \) = variance-covariance matrix among the factors \((T_j \text{ and } e_{ij})\),

and \( \Lambda \) = 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 $\Lambda$ and $\Phi$ 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 $\Sigma$ 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:

\[
\begin{align*}
X_{11} &= b_{11} T_1 + e_{11}, \\
X_{12} &= b_{12} T_2 + e_{12}, \\
X_{21} &= b_{21} T_1 + e_{21}, \\
X_{22} &= b_{22} T_2 + e_{22}, \\
X_{31} &= b_{31} T_1 + e_{31}, \\
X_{32} &= b_{32} T_2 + e_{32},
\end{align*}
\]

and

\[
\begin{align*}
X_{31} = b_{31} T_1 + e_{31}, \\
X_{32} = b_{32} T_2 + e_{32}.
\end{align*}
\]
The vector of factors is \([T_1, T_2, e_{11}, e_{12}, e_{21}, e_{22}, e_{31}, e_{32}]\), and the \(\Lambda\) and \(\Phi\) matrices in equation (2) have the following form:

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

\[
\Phi = \begin{bmatrix}
  \phi_0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
  0 & 0 & \phi_1 & 0 & 0 & 0 & 0 & 0 \\
  0 & 0 & 0 & \phi_2 & 0 & 0 & 0 & 0 \\
  0 & 0 & 0 & 0 & \phi_3 & 0 & 0 & 0 \\
  0 & 0 & 0 & 0 & 0 & \phi_4 & 0 & 0 \\
  0 & 0 & 0 & 0 & 0 & 0 & \phi_5 & 0 \\
  0 & 0 & 0 & 0 & 0 & 0 & 0 & \phi_6 \\
\end{bmatrix}
\]

where the \(b_{ij}\) and the \(\phi_j\) elements are parameters to be estimated.

The specifications for \(\phi\) 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 \( \{ \phi_1, \phi_3, \phi_5, \phi_7, \phi_9 \} \) and three intrajudge error covariances \( \{ \phi_2, \phi_5, \phi_9 \} \) 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 \( \Sigma \) for the above specifications.

The diagonal entries of \( \Sigma \) are variances and are given by:

\[
\text{V}(X_{ij}) = b_{ij}^2 + \text{V}(e_{ij}).
\]  

(3)

The \( \text{V}(e_{ij}) \) are equal to particular diagonal elements in \( \Phi \); for example,

\[
\text{V}(e_{11}) = \phi_1, \quad \text{and} \quad \text{V}(e_{21}) = \phi_3.
\]

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

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

Given three judges it follows that

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

(4)

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

and

\[
b_{3j}^2 = \frac{C(X_{1j}X_{3j}) \cdot C(X_{2j}X_{3j})}{C(X_{1j}X_{3j})}.
\]
Since the $b_{ij}^2$ can be expressed in terms of the elements in $\Sigma$ 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:

$$C(X_{i1}X_{12}) = b_{i1}b_{12}C(T_1T_2) + C(e_{i1}e_{12}),$$

where the $C(e_{i1}e_{12})$ are equal to $\phi_2$, $\phi_5$ and $\phi_8$ for $i$ equal 1, 2 and 3, respectively. Since $b_{i1}$, $b_{12}$, and $C(T_1T_2)$ were shown above to be identifiable, it follows from these equations that the $C(e_{i1}e_{12})$ are identified.

This model has $6 \times 7 / 2 = 21$ unique elements in $\Sigma$ and 16 model parameters (6 in $\Lambda$, 10 in $\Phi$) 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 $\Phi$ matrix. This formulation is convenient for investigating the question of identification as was done above. For purposes of estimation and interpretation, however,
\[ \lambda_1 = \sqrt{\phi_1}, \]
\[ \lambda_2 = \sqrt{\phi_3}, \]
\[ \lambda_3 = \sqrt{\phi_4}, \]
\[ \lambda_4 = \sqrt{\phi_6}, \]
\[ \lambda_5 = \sqrt{\phi_7}, \]
\[ \text{and } \lambda_6 = \sqrt{\phi_9}. \]

The \( \phi \) s are obtained from the \( \phi \) 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 \( \Lambda \) and \( \phi \) that is used in the empirical example presented below.

### III. Empirical Example

Joreskog's (1970) general method for analysis of covariance structures and its associated computer program (Joreskog, Gruvaeus, & van Thillo, 1976) were used for estimation. An earlier program for restricted maximum likelihood factor analysis (Joreskog & Gruvaeus, 1967) or a more recent program for estimating a linear structural equation system (Joreskog & 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 \([T_1 = \text{Dealing with people} \text{ and } T_2 = \text{Responsibility/Autonomy} \text{].} \)
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 $A$ and $\phi$ are given in Table 2.

Table 1. The Observed Correlation Matrix

<table>
<thead>
<tr>
<th></th>
<th>$X_{11}$</th>
<th>$X_{12}$</th>
<th>$X_{21}$</th>
<th>$X_{22}$</th>
<th>$X_{31}$</th>
<th>$X_{32}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_{11}$</td>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$X_{12}$</td>
<td>.5851</td>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$X_{21}$</td>
<td>.2462</td>
<td>.1218</td>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$X_{22}$</td>
<td>.4110</td>
<td>.5360</td>
<td>.2709</td>
<td>1.0000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$X_{31}$</td>
<td>.3823</td>
<td>.2946</td>
<td>.2033</td>
<td>.0694</td>
<td>1.0000</td>
<td></td>
</tr>
<tr>
<td>$X_{32}$</td>
<td>.2816</td>
<td>.6114</td>
<td>.1675</td>
<td>.5049</td>
<td>.3314</td>
<td>1.0000</td>
</tr>
</tbody>
</table>
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

\[
\hat{\Lambda} = \begin{bmatrix}
.766 & 0 & .632 & 0 & 0 & 0 & 0 & 0 \\
0 & .792 & 0 & .607 & 0 & 0 & 0 & 0 \\
.331 & 0 & 0 & 0 & .940 & 0 & 0 & 0 \\
0 & .671 & 0 & 0 & 0 & .742 & 0 & 0 \\
.486 & 0 & 0 & 0 & 0 & 0 & .888 & 0 \\
0 & .769 & 0 & 0 & 0 & 0 & 0 & .645 \\
1.000 & .605 & 1.000 & Symmetric \\
0 & 0 & 1.000 & \\
0 & 0 & .513 & 1.000 \\
0 & 0 & 0 & 0 & 1.000 \\
0 & 0 & .177 & .1.000 \\
0 & 0 & 0 & 0 & 0 & 1.000 \\
0 & 0 & 0 & 0 & .265 & 1.000
\end{bmatrix}
\]
A crucial part of the output is the estimated value of $\Sigma$ 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 residuals of $S - \Sigma$ are given in Table 3.

Table 3. Residuals $S - \Sigma$

<table>
<thead>
<tr>
<th></th>
<th>$X_{11}$</th>
<th>$X_{12}$</th>
<th>$X_{21}$</th>
<th>$X_{22}$</th>
<th>$X_{31}$</th>
<th>$X_{32}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_{11}$</td>
<td>0.013</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$X_{12}$</td>
<td>0.021</td>
<td>0.004</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$X_{21}$</td>
<td>-0.007</td>
<td>-0.037</td>
<td>0.008</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$X_{22}$</td>
<td>0.100</td>
<td>0.005</td>
<td>0.013</td>
<td>-0.001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$X_{31}$</td>
<td>0.010</td>
<td>0.062</td>
<td>0.043</td>
<td>-0.128</td>
<td>-0.024</td>
<td></td>
</tr>
<tr>
<td>$X_{32}$</td>
<td>-0.075</td>
<td>0.003</td>
<td>0.014</td>
<td>-0.011</td>
<td>-0.046</td>
<td>-0.007</td>
</tr>
</tbody>
</table>
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, Joreskog, & 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 \( T_1 \):

\[
\begin{align*}
    b_{11}^2 &= \frac{C(X_{11}X_{21}) C(X_{11}X_{31})}{C(X_{21}X_{31})}, \\
    b_{21}^2 &= \frac{C(X_{11}X_{21}) C(X_{21}X_{31})}{C(X_{11}X_{31})}, \quad \text{and} \\
    b_{31}^2 &= \frac{C(X_{11}X_{31}) C(X_{21}X_{31})}{C(X_{11}X_{21})}.
\end{align*}
\]

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:
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.


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.