DOCUMENT RESUME

TM 002 301

ED 070 777

AUTHOR	Werts, Charles E.; Linn, Robert L.
TITLE	A Review and Synthesis of Educational Measurement Procedures for Studying Growth with the Purpose of Specifying the Appropriate Applications for These Procedures. Final Report.
INSTITUTION	Educational Testing Service, Princeton, N.J.
SPONS AGENCY	Office of Education (DHEW), Washington, D.C. Bureau of Research.
BUREAU NO	0-0352B
PUB DATE	Jun 72
GRANT	OEG-2-700033(509)
NOTE	174p.
EDRS PRICE	MF-$0.65 HC-$6.58
DESCRIPTORS	*Academic Achievement; Analysis of Covariance; Critical Path Method; Educational Research; Factor Analysis; Literature Reviews; *Mathematical Models; *Psychometrics; Research Methodology; *Statistics; Technical Reports
IDENTIFIERS	Joreskog (K G); *Measurement Errors

ABSTRACT

The objective of this study was to review and integrate the various methodologies used in the study of individual growth (especially academic growth). This was accomplished by means of Joreskog's general model for the analysis of covariance structures, i.e., each of the disparate methodologies available from the literature was shown to be a special case of Joreskog's general model. Two general considerations enter into the study of growth and its determinants: (a) making provision for errors of measurement, and (b) constructing a model which relates growth to its determinants in a causally meaningful way. Errors of measurement typically involve questions about the reliability and/or validity of measures, i.e., only indirect measures of the desired variable (construct) are available. Multiple measures of each construct would appear necessary to deal with measurement errors in a quantitative manner. For this purpose the multitrait/multimethod approach devised by Campbell and Fiske (1959) is a useful approach since in principle it allows for correlated errors of measurement. Because the Campbell-Fiske approach does not specify the exact relationships between observed variables and constructs, a factor analytic formulation of their approach was used in order to summarize various approaches to measurement error. The constructs, which represent the growth variable and its determinants, were then interrelated in terms of a linear structural (causal) model. The implications of this model, which itself is a special case of Joreskog's general model, were considered. (Author)

PR-72-9

FINAL REPORT

Project No. 0-0352B

Grant No. OEG-2-700033(509)

A REVIEW AND SYNTHESIS OF EDUCATIONAL MEASUREMENT PROCEDURES FOR STUDYING GROWTH WITH THE PURPOSE OF SPECIFYING THE APPROPRIATE APPLICATIONS FOR THESE PROCEDURES

Charles E. Werts and Robert L. Linn

Educational Testing Service

Princeton, New Jersey 08540

June 1972

U.S. DEPARTMENT OF

HEALTH, EDUCATION, AND WELFARE

Office of Education

Bureau of Research

Final Report

Project No. 0-0352B

Grant or Contract No. OEG-2-700033(509)

A REVIEW AND SYNTHESIS OF EDUCATIONAL MEASUREMENT PROCEDURES FOR STUDYING GROWTH WITH THE PURPOSE OF SPECIFYING THE APPROPRIATE APPLICATIONS FOR THESE PROCEDURES

Charles E. Werts and Robert L. Linn

Educational Testing Service

Princeton, New Jersey 08540

June 1972

The research reported herein was performed pursuant to a grant with the Office 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 Office of Education position or policy.

U.S. DEPARTMENT OF

HEALTH, EDUCATION, AND WELFARE

Office of Education

Bureau of Research

Table of Contents

Acknowledgments

The authors are indebted to Dr. Karl Jöreskog, Uppsala University, Sweden, who assisted us in interpreting the growth methodology literature in terms of his general model for the analysis of covariance structures. Dr. Hubert Blalock, University of Washington, Dr. Arthur Goldberger, University of Wisconsin, and Dr. Frederic Lord, Educational Testing Service, also provided valuable consultation.

I. Summary

The objective of this study was to review and integrate the various methodologies used in the study of individual growth (especially academic growth). This was accomplished by means of Jöreskog's general model for the analysis of covariance structures, i.e., each of the disparate methodologies available from the literature was shown to be a special case of Jöreskog's general model. Two general considerations enter into the study of growth and its determinants: (a) making provision for errors of measurement and (b) constructing a model which relates growth to its determinants in a causally meaningful way. Errors of measurement typically involve questions about the reliability and/or validity of measures, i.e., only indirect measures of the desired variable (construct) are available. Multiple measures of each construct would appear necessary to deal with measurement errors in a quantitative manner. For this purpose the multitrait-multimethod approach devised by Campbell and Fiske (1959) is a useful approach since in principle it allows for correlated errors of measurement. Because the Campbell-Fiske approach does not specify the exact relationships between observed variables and constructs, a factor analytic formulation of their approach was used in order to summarize various approaches to measurement error. The constructs, which represent the growth variable and its determinants, were then interrelated in terms of a linear structural (causal) model. The implications of this model, which itself is a special case of Jöreskog's general model, were considered.

II. Introduction

This project was a review and synthesis of educational measurement methodologies for studying growth. To this end the initial phases consisted of a review of relevant literature in econometrics, psychometrics, statistics and sociometry. Some of the concepts which developed from this review seemed worthy of immediate dissemination via formal and informal publication media. In particular the following articles commented on separate aspects of our review:

Werts, Charles E., Jöreskog, Karl G., & Linn, Robert L. Comment on "The estimation of measurement error in panel data." American Sociological Review, 1971, 36, 110-113.

Werts, Charles E., & Linn, Robert L. Comment on Boyle's "Path Analysis and Ordinal Data." American Journal of Sociology, 1971, 76, 1109-1112.

Werts, Charles E., & Linn, Robert L. Errata to the Werts-Linn Comments on Boyle's "Path Analysis and Ordinal Data." American Journal of Sociology, 1972, in press.

Werts, Charles E., Linn, Robert L., & Jöreskog, Karl G. Another perspective on "Linear regression, structural relations, and measurement error." Educational and Psychological Measurement, in press.

Werts, Charles E., Linn, Robert L., & Jöreskog, Karl G. A congeneric model for platonic true scores. Research Bulletin 71-22, Educational Testing Service, Princeton, New Jersey, May 1971. Also in Educational and Psychological Measurement, in press.

Werts, Charles E., & Linn, Robert L. Estimating true scores using group membership. Educational and Psychological Measurement, in press.

Linn, Robert L., & Werts, Charles E. Errors of inference due to errors of measurement. Research Bulletin 71-7, Educational Testing Service, Princeton, New Jersey, February 1971. Also in Educational and Psychological Measurement, in press.

Werts, Charles E., Jöreskog, Karl G., & Linn, Robert L. Identification and estimation in path analysis with unmeasured variables. Research Bulletin 71-39, Educational Testing Service, Princeton, New Jersey, June 1971. Also in American Journal of Sociology, in press.

Werts, Charles E., Linn, Robert L., & Jöreskog, K. G. Intraclass reliability estimates: testing structural assumptions. Educational and Psychological Measurement, in press.

Copies of these articles are included in the Appendix. Those aspects directly relevant to the project goal are treated in the review sections which follow.

For heuristic purposes the review and synthesis of the literature has been treated in two parts. The first part (Sec. III) labelled "Quantifying Unmeasured Variables" treats the general methodological considerations relevant to growth studies and a wide variety of the problems involving errors of measurement and causal analyses. This part will appear in a new book, Theories and strategies of measurement in the social sciences, H. M. Blalock, editor. Blalock's books are widely used in the social sciences as textbooks.

The second part of our review (Sec. IV) labelled "A multitrait-multimethod model for studying growth" reviews various psychometric formulations specifically relevant to growth studies and formally treats them as a special case of Jöreskog's general model for the analysis of covariance. Implications for factor analytic studies of growth data and for studies of the determinants of growth are detailed. This part will appear in Educational and Psychological Measurement and has been released in preliminary form using the Educational Testing Service Research Bulletin series.

III. General Methodological Considerations: Quantifying Unmeasured Variables

Social scientists frequently wish to make inferences about the "effects" of hypothetical constructs which are not directly measured, e.g., only the symptoms, antecedents, and/or consequences of the construct may be measurable. In recent years a variety of statistical procedures have been introduced to help quantify the relationships among observed variables and constructs in an attempt to increase the rigor and validity of such inferences. The purpose of this essay is to introduce the various concepts and to consider the numerous assumptions involved in these procedures so that the user will be aware of analytical potentials and limitations.

1. Validity

A basic concept in the discussion of indirectly measured concepts is that of validity. This refers to the relationship between an observed variable (X) and the unmeasured construct (Y). We shall discuss models in which it is assumed that the relationship is linear, i.e.,

$$X = bY + I + e \tag{1}$$

where b is the slope of the regression of X on Y, I is the intercept of this regression line, and e is a residual which is taken to be independent of Y. Econometricians (e.g., Goldberger, 1970) typically specify $b = 0$, $I = 0$, and e is labelled a disturbance instead of the psychometric term errors of measurement. Despite the crucial importance of this linear relationship, it is seldom that data analysts substantively justify this assumption. For example, ability and achievement test scores are generally assumed to have a linear relationship with their underlying true scores, however Carver (1969) has persuasively argued that there is a curvilinear relationship between knowledge (the construct) and test scores in classroom learning, i.e., more knowledge is required to increase the test score one point at the high end of the scale. When psychologists use the term validity coefficient they are usually referring to the correlation (i.e., R_{XY}) between the observed variable and the construct (i.e., true score) assuming the residuals of X on Y to be independent of Y (Guilford, 1954, Chap. 14). As long as consideration is limited to a single variable X and a single construct Y the linear relationship is not a real limitation, unless an added constraint such as equal intervals is added, because the Y could be transformed to yield a linear relationship with X. With two X's for a single construct the limitation becomes a real one.

It is useful to distinguish between the terms reliability and validity. A traditional test theorist will typically consider the correlation between parallel forms (X_1 and X_2) of a test to be the reliability coefficient. As illustrated in Fig. 1.a, the model here is $X_1 = b_1Y + I_1 + e_1$ and $X_2 = b_2Y + I_2 + e_2$ where e_1 and e_2 are assumed independent of each other and of Y ; which implies that $R_{e_1e_2} = R_{e_1Y} = R_{e_2Y} = 0$. Test forms are said to be parallel when the variances of e_1 and e_2 are equal (i.e., $V_{e_1} = V_{e_2}$) , $b_1 = b_2$ and $I_1 = I_2$. It follows that for parallel forms the correlation between the observed measures will equal the square root of the correlation of either measure with the construct, i.e., $\sqrt{R_{X_1Y}} = \sqrt{R_{X_2Y}} = R_{X_1X_2}$ =reliability coefficient. If the variable which is being measured by the parallel forms (i.e., Y) is itself a symptom of another construct (e.g., Z) then new assumptions must be made, e.g., $Y = bZ + \mu$ where μ is independent of Z , e_1 and e_2 as shown in Fig. 1b. In this case the correlations between parallel observed measures and Z are $R_{X_1Z} = R_{X_2Z} = R_{YZ} R_{X_1Y} = R_{YZ} R_{X_2Y} = R_{YZ} R^2_{X_1X_2}$. In this model the X_1 on Z residuals have the form $(X_i - b_i bZ) = b_i\mu + e_i$ and the covariance between the X_1 and X_2 on Z residuals will equal $b_1 b_2 V_\mu$. Therefore these residuals are in general correlated and $R_{X_1X_2} \neq \sqrt{R_{X_1Z}}$ or $\sqrt{R_{X_2Z}}$. R_{X_1Z} and R_{X_2Z} cannot be estimated, however $R^2_{X_1X_2}$ is the upper limit for these correlations, i.e., reliability sets an upper bound on validity. For illustrative purposes consider the problem of measuring achievement in mathematics for 9th grade students in city A. Two (or more) parallel forms of widely used mathematics tests, standardized on national samples, can be readily obtained and administered. These forms typically have very similar item formats, the items differing mainly with respect to the numbers inserted in the problems. Because these tests cater to a wide variety of schools the items necessarily cover material which is common to most curricula at this level. Insofar as the curriculum in city A has special emphasis, not generally taught elsewhere, the nationally standardized tests will be partly irrelevant (i.e., invalid) to city A. The parallel forms would correspond to X_1 and X_2 in Fig. 1.b, the variable Y would represent achievement on generally taught problems, and Z would be the achievement of students in city A. If the discrepancy between Y and Z is very great, as inferred from curricular differences, then city A could build equivalent forms which more precisely cover their coursework, which might then correspond to the model in Fig. 1.c. It is always necessary

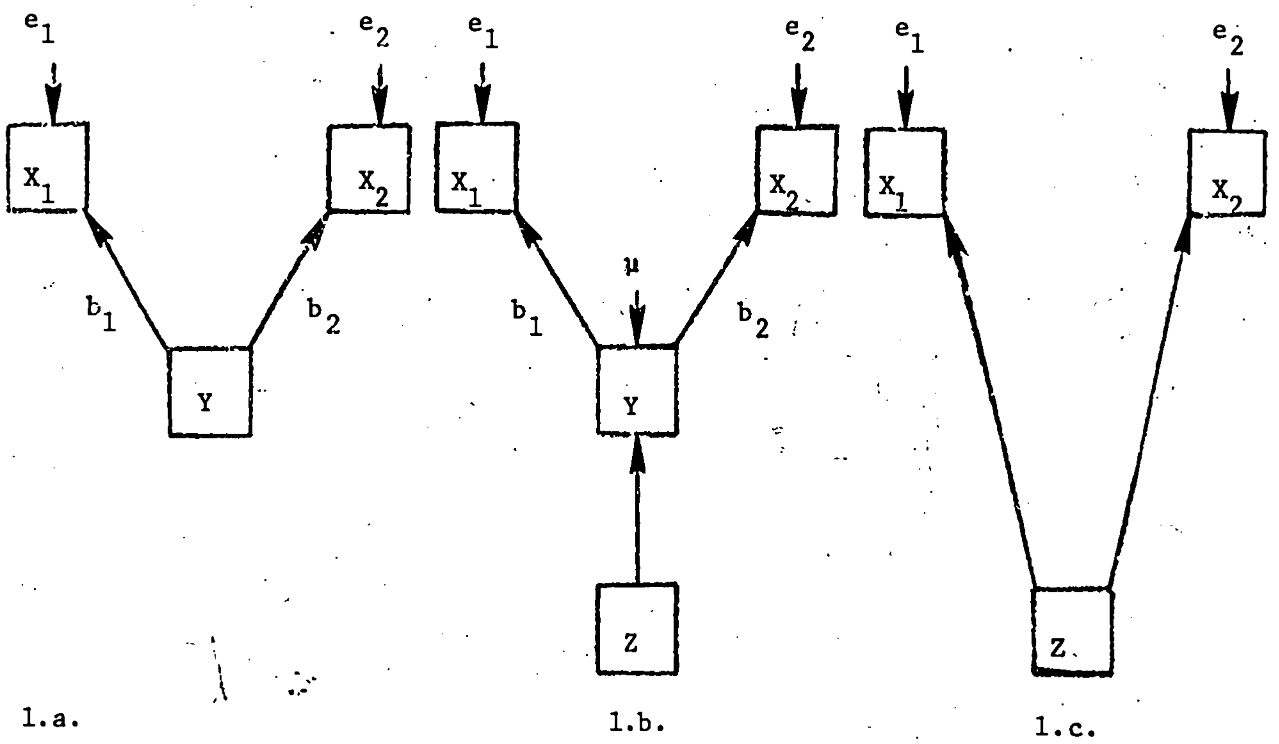


Fig. 1. Reliability and validity models.

for the researcher to examine test materials in order to see how well the construct being measured by that test corresponds to the construct relevant to the research project. In many cases he may decide to use two measures of a construct with very different types of item formats in order to obtain a model like Fig. 1.c, i.e., the very similarity of item formats may give the scores some covariation which does not represent association due to the underlying construct to be measured (as in Fig. 1.b).

Instead of validity coefficients, factor analysts (e.g., Harman, 1967) refer to factor loadings. A factor loading is the regression weight of an observed score on a factor (viz., construct). The models in Fig. 1.a and 1.c correspond to a single factor model and the standardized factor loading is equal to the correlation of the observed score with the factor like the corresponding reliability and validity coefficients. If there were more than one factor, but these factors were uncorrelated as in an orthogonal solution, then the standardized factor loading would still equal the correlation. In the case of correlated factors as in an cblique solution, the standardized factor loadings are standardized partial regression weights which are called pach coefficients by path analysts (e.g., Duncan, 1966; Wright, 1934).

The regression weight in Equation (1) basically states the relationship between the units of measurement of the observed variable and that of the construct. A weight equal to unity corresponds to the assumption that the observed measure and the construct have the same units of measurement. Psychological test theorists and econometricians usually make this assumption, whereas path (Blalock, 1969; Costner, 1969) and factor analysts commonly assign the factor a variance of unity (i.e., $V_Y = 1$). As shall be noted later, this assumption creates no difficulty until the problem involves multiple measures of a construct and/or growth along the same dimension over time (Werts, Jöreskog, & Linn, 1972).

2. Multiple Measures of a Single Construct

Although econometricians rarely are concerned with multiple measures of a construct, test theorists and path and factor analysts have written extensively on this topic. Much of modern test theory (Lord & Novick, 1968) is derived assuming at least two tau equivalent measures of the underlying true score (i.e., construct). Tau equivalent measures (e.g., X_1 and X_2) are those in which the observed on true regression weights are unity (i.e., $b_1 = b_2 = 1$), the intercepts are equal (i.e., $I_1 = I_2$) and the errors of measurement are independent of each other and of the true score. Essentially tau equivalent measures are the same except that $I_1 \neq I_2$. In contrast to the parallel forms assumptions discussed previously, the error variances are not assumed equal (i.e., $V_{e_1} \neq V_{e_2}$) for tau equivalent or essentially tau equivalent measures, which means that the tests may have different reliabilities (i.e., differing error variances). Since by assumption $X_1 = Y + I_1 + e_1$ and $X_2 = Y + I_2 + e_2$, the covariance $C_{X_1X_2} = V_Y$, i.e., the covariance between the observed scores is equal to the variance of the true scores. The true variance divided by the observed variance (e.g., V_{X_i}) for a test yields the reliability, i.e., $V_Y \div V_{X_i}$.

Essentially tau equivalent and tau equivalent measures assume that the observed measures of the construct have the same units of measurement. When measuring different symptoms or indicators of an underlying construct it is quite common to have different units, e.g., income and occupation as indicators of socioeconomic status typically are measured in different units. In this case the unit of the construct is arbitrary and is usually fixed by assigning a variance of unity, although it is also possible to identify the unit of one of the measures with that of the construct by specifying the corresponding regression weight to be unity. Jöreskog (1971) calls the various measures of the construct <u>congeneric</u> measures ($b_1 \neq b_2 \neq b_i$), whereas factor analysts would say that a single factor structure has been assumed. In each case the errors or residuals are assumed independent of each other and of the construct.

3. Identification

The concept of <u>identification</u> is crucial to any comparison of methods. Mathematicians and econometricians (e.g., Fisher, 1966) have long been interested in developing procedures for dealing with identification problems. Whereas true score theorists and path analysts usually attempt to build identified models, the majority of factor analysts have dealt with highly underidentified models. Although in principle sociologists were exposed to the identification issue in relation to latent structure analysis (e.g., Lazarsfeld, 1950), the recent papers on this subject by path analysts (e.g., Boudon, 1965; Blalock, 1966) have probably had a wider impact. The term <u>identifiable</u> will be used here in the sense defined by Fisher (1966, p. 25): "We shall speak of that equation as identifiable (or identified) if there exists some combination of prior and posterior information which will enable us to distinguish its parameters from those of any other equation in the same form."

To illustrate the identification problem let us consider a single factor model from the perspective of path analysis (Costner, 1969). Suppose we are given four observed measures (X_1, X_2, X_3, X_4) of the factor (Y). The single factor model specifies that $X_i = b_iY + I_i + e_i$ where all e_i are independent of each other and of Y.

The model is depicted in Figure 2 using path analysis notation, i.e., when variables are independent no arrows connect them. To obtain the <u>expected</u> covariances (C_{ij}) between two observed measures $(X_i$ and $X_j)$ we would multiply the corresponding pair of equations to obtain:

$$C_{ij} = b_i b_j V_Y \ , \tag{2}$$

(3) and $$V_{X_i} = b_i^2 V_Y + V_{e_i} \ .$$

The term <u>expected</u> refers to the value of a parameter to be expected in a model without <u>sampling</u> or <u>model specification</u> errors. <u>Specification errors</u> refer to the incorrect choice of a statistical model (Theil, 1957). It is convenient to arrange the expected variances and covariances given by equations (2) and (3) into an expected <u>variance-covariance matrix</u> (Σ) , e.g., in the four variable case:

$$\Sigma = \begin{bmatrix} V_1 & C_{12} & C_{13} & C_{14} \\ C_{12} & V_2 & C_{23} & C_{24} \\ C_{13} & C_{23} & V_3 & C_{34} \\ C_{14} & C_{24} & C_{34} & V_4 \end{bmatrix}$$

To see if this model is identified, the path analyst (e.g., Costner, 1969) typically would standardize all variables $(V_{X_1} = V_{X_2} = V_{X_3} = V_{X_4} = V_Y = 1)$ and then derive the equations for each expected correlation (R_{ij}) in terms of the path coefficients (b_i^*) of the model, e.g.,

$$R_{12} = b_1^* b_2^* \ ,$$

$$R_{13} = b_1^* b_3^* \ ,$$

$$R_{14} = b_1^* b_4^* \ ,$$

$$R_{23} = b_2^* b_3^* \ ,$$

$$R_{24} = b_2^* b_4^* \ ,$$

and $$R_{34} = b_3^* b_4^* \ .$$

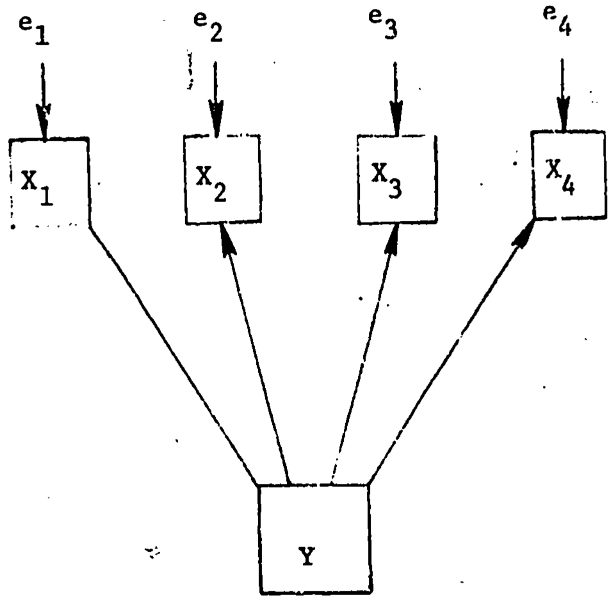


Fig. 2. A single factor model.

Using any three measures $(X_i, X_j$, and $X_k)$ it is possible to solve for the unknown $(b_i^*)^2 = (R_{ij}R_{ik}) \div R_{jk}$. Thus all parameters (b_i^*) are identified, in the sense that each parameter may be stated as a function of potentially observable information. The actually observed sample variances and covariances could also be arranged in a matrix (S). The observed matrix (S) may differ from the expected matrix (Σ) because of sampling and specification errors. The model is usually judged to be incorrect if Σ and S differ very much, i.e., when the observed data does not fit the model. Quite sophisticated techniques are now available to obtain parameter estimates which minimize in some sense the difference between the observed matrix and the expected matrix computed from the parameter estimates (Hauser & Goldberger, 1970; Jöreskog, 1970).

The equations relating the expected correlations (R_{ij}) to the model parameters (b_i^*) are called path equations by path analysts. When the parameters are identified by these equations, a model is called just identified if the number of observable quantities (R_{ij}) equals the number of unknown parameters (b_i^*) in the path equations and overidentified if the observables exceed the parameters. If the number of unknown parameters exceeds the number of observables, then the model is underidentified even though a subset of the parameters may be identified.

Jöreskog labels models which are overidentified as confirmatory. In confirmatory factor studies the experimenter has already obtained a ertain amount of knowledge about the variables measured and therefore is in a position to formulate a model which is to be tested for fit to data. Most factor analysts deal with highly underidentified models; exploratory factor procedures being used to suggest an appropriate number of factors to use and a preliminary interpretation of the data. In contrast, econometricians, path analysts, and classical test theorists usually deal with identified models which reflect substantive theoretical considerations. It is logically possible for the model suggested by exploratory procedures to be identified, but factor analysts have typically not examined this question because their main interest is in fit, not in identifiability.

4. Multifactor Models

Let us consider a simple two factor $(Y_1$ and $Y_2)$ model (Fig. 3) in which there is only one observed measure $(X_1$ and $X_2)$ of each factor, i.e., $X_1 = b_1Y_1 + I_1 + e_1$ and $X_2 = b_2Y_2 + I_2 + e_2$

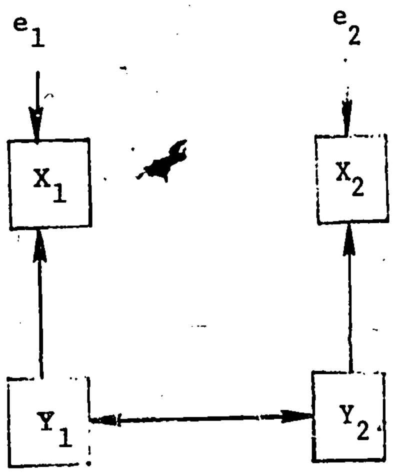


Fig. 3. A simple two factor model.

where e_1 and e_2 are independent of each other and Y_1 and Y_2. When all variables are standardized there is one observed correlation (R_{12}) and three unknown correlations ($R_{X_1Y_1} = b_1^*$, $R_{Y_1Y_2}$, and $R_{X_2Y_2} = b_2^*$) among variables (i.e., the model is underidentified) and $R_{12} = R_{X_1Y_1} R_{Y_1Y_2} R_{X_2Y_2}$. Psychometricians call the correlation between the factors ($R_{Y_1Y_2}$) the unattenuated correlation. In the case of tests, the publisher usually provides test reliabilities (labelled R_{11} and R_{22}) which in this model might be used to estimate (denoted by "^") the square of the correlation with the appropriate factor, i.e., R_{11} $\hat{R}^2_{X_1Y_1}$ and $R_{22} = \hat{R}^2_{X_2Y_2}$. Given these reliabilities we may estimate the correlation between factors as:

$$\hat{R}_{Y_1Y_2} = R_{12} \div \sqrt{R_{11} R_{22}} .$$

This procedure is called correcting for attenuation.

A. Exact Functional Relationship Among Factors

Statisticians (e.g., Kendall & Stuart, 1961) and econometricians (e.g., Johnston, 1963, Chap. 6) have been interested in the variation of the Fig. 3 model in which $b_1 = b_2 = 1$ and the factors have an exact functional relationship, i.e., $Y_2 = I + BY_1$ and $R_{Y_1Y_2} = 1$. It might for example be hypothesized that in a class of equally intelligent and motivated students, the amount they will learn in a math course (Y_2) will be directly proportional to their relevant mathematics skills (Y_1) at the beginning of the course because, e.g., those who know more are better able to understand the teacher.

Neglecting variable means, since there are three unknown parameters (b_1, b_2, B) and only one observed correlation (R_{12}), this model is underidentified. Isaac (1970) reviews the estimating formulae for the case in which the error variances V_{e_1} and/or V_{e_2} or their ratio $V_{e_1} \div V_{e_2}$ are known.

B. Stochastic Components

Johnston (1963, p. 148) notes that the exact functional relationship model discussed above "hardly seems appropriate for econometric work, since if it were true, all points would be exactly on a straight line. A stochastic component of behavior would seem an essential in economics." This comment probably applies to all the social sciences in which it is generally necessary in linear structural models to assume that all the other <u>unmeasured</u> variables influencing a variable of interest are independent of the influences that are measured (Blalock, 1967). It seems most unlikely, for example, that there are not other disturbing factors which will influence mathematics achievement.

Adding a stochastic disturbance term, μ , representing these other variables, the equation between the factors becomes $Y_2 = I + \beta Y_1 + \mu$ where μ is independent of Y_1 and $b_1 = b_2 = 1$. The analysis of this stochastic model is discussed by Johnston (1963, Chap. 6). One approach assumes that the error variances V_{e_1} and V_{e_2} are known, which is equivalent to the psychometrician's approach since knowing the error variances the reliabilities can be computed, i.e., $R_{ii} = V_{Y_i}/V_{X_i}$ where $V_{Y_i} = V_{X_i} - V_{e_i}$. Because $R_{Y_1Y_2}$ is identified by the formula for attenuation, it follows that β and therefore V_μ are also identified, i.e., $\beta = R_{Y_1Y_2}\sqrt{V_{Y_2} \div V_{Y_1}}$ and $V_\mu = V_{Y_2} - \beta^2 V_{Y_1}$. The difficulty with this approach lies in the problem of obtaining reasonable estimates of the error variances. Even when reliabilities are given as in the case of many published tests, these figures may be erroneous to an unknown degree for the particular subpopulation being tested.

Another approach is the use of <u>instrumental</u> variables, i.e., in this case, a variable (Z) which is independent of both the errors e_1 and e_2 . In this case the regression weight β may be estimated as $\hat{\beta} = \text{cov}\ (Y_2Z) \div \text{cov}\ (Y_1Z)$. It may be shown that the reliability coefficient for X_1 is $\hat{R}_{11} = R_{X_1Z}R_{X_1X_2} \div R_{X_2Z}$, which from the previous section can be seen as the solution for the squared factor loading in the single factor model in which X_1 , X_2 , and Z are <u>congeneric</u> measures $\left[\hat{R}^2_{X_2Y_1} = R_{X_1X_2}R_{X_2Z} \div R_{X_1Z}\right.$ and $\left.\hat{R}^2_{Y_1Z} = R_{X_1Z}R_{X_2Z} \div R_{X_1X_2}\right]$. Further analysis would show that V_{e_2} and V_μ are not identified. The basic problem in use of instrumental variables is that we are seldom in a position to check whether this variable is in fact independent of errors, yet the estimates are likely to be highly dependent on which such variable is selected (Blalock, Wells, & Carter, 1970). The same problem plagues the use of the congeneric model since it is seldom obvious exactly which observed measures really are indicators of the same underlying trait assuming independent errors. It is interesting to note that in these models an instrumental variable substitutes for a congeneric measure, i.e., what is needed is a third measure which is independent of the errors in the other two variables. For illustrative purposes consider the problem of measuring differential student math achievement given the scores from two different nationally distributed objective exams, one perhaps using a problem format and another a multiple choice format; whose validities for the curriculum of interest are unknown. A third congeneric measure might well be the course grades given by the teacher. The logic here is that these should all be tapping the achievement dimension but to differing degrees and there is no a priori reason to believe that errors of measurement among these measures are correlated since very different formats are involved. Sometimes, however, achievement tests are given in batteries such that the needed third measure might be in another content area. For example, English achievement scores might be available. It is unlikely that this test is correlated with errors of measurement on the two objective math tests and this could therefore serve as an instrumental variable.

C. Model with Multiple Indicators

Economists (e.g., Goldberger, 1970) and sociologists (e.g., Blalock, 1969; Costner, 1969) rarely have the data to estimate reliability from independent sources, whereas psychometricians and factor analysts (at least implicitly) frequently do so. A traditional technique of this type used by psychometricians is the split half procedure (e.g., Guilford, 1954, p. 377). The items on a test are split in half (e.g., odd items assigned to one-half and even to the other) and the correlation between the halves used to estimate the reliability of the whole test, assuming that the halves are equivalent measures. Various formulae are used to adjust for the fact that the halves are not as long as the whole test and therefore not as reliable (Guilford, 1954, Chap. 14). These reliability estimates may then be used to estimate the unattenuated correlation between two tests, i.e., the correlation between the two true factors underlying the observed measures.

The logic of the split half approach is worth further study. Changing to a double subscript for each observed measure (X_{ij}) where j refers to the j^{th} construct (Y_j) and i to the i^{th} indicator of the j^{th} construct; then in the split half procedure the equations are:

$$X_{11} = b_{11}Y_1 + I_{11} + e_{11} ,$$

$$X_{21} = b_{21}Y_1 + I_{21} + e_{21} ,$$

$$X_{12} = b_{12}Y_2 + I_{12} + e_{12} , \quad \text{and}$$

$$X_{22} = b_{22}Y_2 + I_{22} + e_{22} .$$

Using path analytic procedure we find that:

$$R_{X_{11}X_{21}} = b^*_{11}b^*_{21} ,$$

$$R_{X_{11}X_{12}} = b^*_{11}R_{Y_1Y_2}b^*_{12} ,$$

$$R_{X_{11}X_{22}} = b^*_{11}R_{Y_1Y_2}b^*_{22} ,$$

$$R_{X_{21}X_{12}} = b^*_{21}R_{Y_1Y_2}b^*_{12} ,$$

$$R_{X_{21}X_{22}} = b^*_{21}R_{Y_1Y_2}b^*_{22} , \quad \text{and}$$

$$R_{X_{12}X_{22}} = b^*_{12}b^*_{22} .$$

Solution of these equations indicates that all the reliabilities (b^*_{ij}) and the unattenuated correlation $(R_{Y_1Y_2})$ are identified without further assumptions. This model is overidentified since there is one more equation than unknown parameters, i.e., there is one degree of overidentification which is equivalent to one degree of freedom in Jöreskog's (1970) general model for the analysis of covariance structures (which may be used for estimation purposes). Because the model is identified we may check to see if it is reasonable to believe that $b^*_{11} = b^*_{21}$, $b^*_{12} = b^*_{22}$, $V_{e_{11}} = V_{e_{21}}$, and $V_{e_{12}} = V_{e_{22}}$ as asserted in the assumption that split halves are parallel (Werts & Linn, 1971). Even without the assumption of parallel halves the model may be tested for fit to the data. As Guilford (1954, p. 377) notes, the difficulty with the odd-even method is that the observed correlation between the splits will generally be too high because of "extra-test determiners contributing positively to the observed correlation." For example, testing conditions and amount of time devoted to each half will be nearly constant for the halves. In contrast the alternate forms method, with at least a day between administrations, introduces a change of conditions which "is more like those changes between administration of two different tests or between test administration and measurement of some criterion in validation" (Guilford, 1954, p. 377). If these other determiners were independent of the true score then in our model these would be equivalent to asserting that the corresponding errors were not in fact independent (e.g., $R_{e_{11}e_{21}} \neq 0$). If this were the case, it is possible that this might be detected as a lack of model fit to the observed data. Psychometricians have various other procedures for estimating whole test reliability from item data (Stanley, 1971), the logic being much like that discussed here except that each item now becomes an observed measure. To the degree that the item data do not fit a single factor model these estimates become difficult to interpret (Werts & Linn, 1970a). Nonetheless, in practice this fit is seldom checked.

5. The Multitrait-Multimethod Approach

The multitrait-multimethod matrix technique (Campbell & Fiske, 1959) has been of considerable interest to psychologists because it provides information on the convergent (confirmation by independent measurement procedures) and discriminant (separation of one trait from another) validity of theoretical constructs (i.e., traits). The problem of measuring mathematics achievement as opposed to achievement in English may be used to illustrate these concepts. To measure math achievement we might use three measures including one "subjective" measure, course grades, and two "objective" measures consisting of a multiple choice and a mathematics reasoning test (perhaps constructed by the publisher of the course material). Despite the differences in format, each measure in principle is simply another demonstration of the student's grasp of the subject matter and should therefore tend to give fairly consistent results. Insofar as the results are indeed consistent, convergent validity is demonstrated. The logic underlying convergent validity is much like that of the congeneric model previously discussed. The emphasis on different methods of measurement represents an attempt to ensure that the correlations among variables as much as possible represent commonality with the underlying trait rather than consistencies due to similarities of testing methods. Thus, use of different methods tends to support the assumption of independent errors required by the congeneric model. Now suppose that English achievement were also obtained from three measures whose format was like that used for math achievement, i.e., course grades, a multiple choice and a reasoning test. Discriminant validity would be demonstrated if it could be shown that the trait (i.e., factor) underlying the math measures were distinctly different from the trait underlying the English measures. According to Campbell and Fiske, convergent validity is demonstrated by at least moderate correlations between different methods measures of the same trait and discriminant validity is shown by a higher correlation between independent efforts (i.e., methods) to measure the same trait than between measures designed to get at different traits using the same method. From our perspective discriminant validity consists of demonstrating that the true correlation between two traits is meaningfully less than unity. Werts and Linn (1970b) have discussed the Campbell-Fiske approach from this perspective. The analytical procedures devised by Campbell and Fiske (1959) are not of interest here because no attempt was made to specify the nature of the relationship between the observed measures and the trait or methods factors. It should be clear from our previous statements that an observed variance-covariance matrix is interpretable only from the perspective of an hypothesized model. Campbell and Fiske's argument that the researcher should obtain measures of a trait which differ as much as possible in measurement technique, in order to improve convergent validity, is very pertinent. From the multitrait-multimethod perspective the typical psychometric approach, which attempts to devise alternate forms with almost identical format, would be criticized as lacking in convergent validity.

A variety of analytical methods have been proposed for multitrait-multimethod data (e.g., Boruch, Larkin, Wolins, & McKinney, 1970), however only Jöreskog's confirmatory factor analytic approach will be considered here. Suppose that it were assumed that each observed measure were a function of only one trait (Y_j) and one method (M_k) factor in a linear fashion, i.e.,

$$X_{jk} = a_{jk}Y_j + b_{jk}M_k + I_{jk} + e_{jk}$$

where X_{jk} = measure reflecting combination of trait j and method k

a_{jk} = regression weight of X_{jk} on trait Y_j , and

b_{jk} = regression weight of X_{jk} on method M_k .

Assume also that all residuals are independent of each other and of all factors. It may be shown that at least three traits and three methods must be used in order for this model to be identified, given that all factors may be <u>oblique</u>, i.e., correlated. To understand the connection with models discussed earlier, consider two different method measures of the same trait, e.g., X_{11} and X_{12} (illustrated in Fig. 4). It can be seen that there are several sources of the observed correlation $R_{X_{11}X_{12}}$, i.e., $R_{X_{11}X_{12}} = a^*_{11}a^*_{12} + a^*_{11}R_{Y_1M_2}b^*_{12} + a^*_{12}R_{Y_1M_1}b^*_{11} + b^*_{11}R_{M_1M_2}b^*_{12}$. If the methods factors were independent of the trait factor, the model would in principle be like a congeneric model with correlated residuals. Such a model has been proposed by Guttman (1953) in relation to obtaining reliability estimates from nonindependent item data. If the methods factors were independent, we would have the

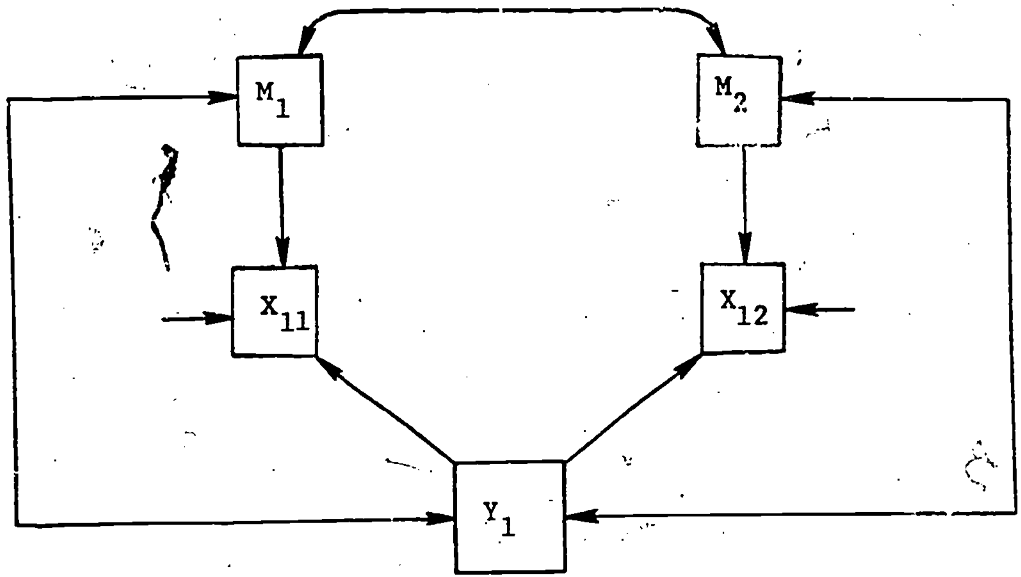


Fig. 4.

congeneric model basic to true score theory. Thus we see that the traditional test theory approaches discussed earlier may be considered the special case of the multitrait-multimethod approach in which methods factors are assumed to be independent of each other and of the trait factors. The notion of reliability as the ratio of true variance to observed variance is only meaningful in the case where errors are independent in this way, i.e., no such neat partitioning of variance is possible in the general multitrait-multimethod approach.

6. Functional Relationships Among Factors

Whereas the econometricians and path analysts postulate functional relationships among the factors, psychologists and factor analysts seldom do so. Both the multitrait-multimethod approach and true score theory focus only on errors of measurement. In part, this situation arises because psychologists are usually taught to avoid making causal inferences from correlations. Sometimes antecedent (i.e., causally prior) variables are statistically controlled in order to insure that a particular correlation is not spurious, however systematic procedures for analyzing sources of a correlation (e.g., path analysis) are viewed with suspicion.

The function of causal hypotheses can be illustrated by an example taken from Werts and Linn (1970b). Suppose there were a linear causal relationship between variables ($Y_2 = BY_1 + \mu$); where Y_2 is measured directly and Y_1 indirectly by two indicators ($X_1 = b_1Y_1 + e_1$, $X_2 = b_2Y_1 + e_2$) . This is a single factor model and $B*$ may be estimated as:

$$(\hat{B}*) = \sqrt{R_{X_1Y_2}R_{X_2Y_2} \div R_{X_1X_2}} \; .$$

For example, if $R_{X_1Y_2} = .20$, $R_{X_2Y_2} = .40$, and $R_{X_1X_2} = .80$ then $\hat{B}* \cong .32$. Most educational psychologists, in their search for school effects, would not even consider the possibility that several measured variables might be indicators of the same underlying construct (i.e., Y_1) and would proceed using the regression equation:

$$Y_2 = b_1X_1 + b_2X_2 + I_3 + e_3 \; ,$$

yielding standardized weights of $b_1^* \cong -.33$ and $b_2^* \cong +.67$. If for example X_1 were proportion of faculty with doctorates and X_2 were number of books per pupil in the library, it might well be supposed that both of these variables are indicators of school affluence (i.e., Y_1) . Certainly the regression procedure, which is typical of school effects studies, would yield no hint of how Y_1 influences Y_2 , i.e., the weights b_1^* and b_2^* are opposite in sign, yet both reflect the same underlying variable. The use of regression equations in this way represents an attempt to avoid theory, finding influences by seeing if a variable increases the percentage of predictable variance in the outcome. It is better to specify the theoretical structure being postulated, so that appropriate analytical procedures may be designed.

A. Growth Studies

Another area where it is important to specify functional relationships is in the study of the determinants of growth. Test theorists have long been concerned with the problem of estimating growth in the presence of errors of measurement (e.g., Harris, 1963). The special feature of this area is that an initial status and a final status are assumed to have identical units of measurement. If the initial status is $X_1 \; b_1Y_1 + I_1 + e_1$ and the final status $X_2 = b_2Y_2 + I_2 + e_1$, then the equal units assumption is equivalent to $b_1 = b_2$. Various procedures (e.g., Cronbach & Furby, 1970) attempt to estimate the true change $Y_2 - Y_1$ from the observed scores and known reliability coefficients for the initial and final measures. From these data a measure of the reliability of differences, i.e., the correlation of the observed difference $X_2 - X_1$ with the true difference $Y_2 - Y_1$ may be obtained. It was originally thought that if the reliability of differences was low then our ability to estimate true change would be low; however, Cronbach and Furby (1970) and Werts and Linn (1970a) have demonstrated the use of information on other variables to help estimate change. The logic of this approach is an extension of the rationale enunciated earlier with regards to instrumental variables, i.e., both causes, effects, and other correlates of growth carry information which can be used to estimate model parameters and therefore to improve estimates of factor scores.

Several educational researchers (Bloom, 1964; Thorndike, 1966) have been concerned with the determinants of $(Y_2 - Y_1)$ and in essence have argued that if the initial status (Y_1) is uncorrelated with gain $(Y_2 - Y_1)$ then the determinants of change during this time interval are different from those which produced the initial level of competence (Y_1). No such conclusion is warranted (Werts, Jöreskog, & Linn, 1972) since without including in the functional model various determinants of growth, it is impossible to make any statements about the effect of these determinants. As the path analysts have so frequently shown, no correlation, even zero, is interpretable in a causal sense except in the framework of a causal model. It is quite possible because of counterbalancing influences, for $Y_2 - Y_1$ to be uncorrelated with Y_1 and yet initial status may influence gain either positively or negatively.

An important feature of growth studies is that the variance of the initial and final status factor (Y_1 and Y_2) is identified by the scaling assumption $b_1 = b_2$. For convenience test theorists usually assign the value $b_1 = b_2 = 1$, i.e., that the factors have the same units as the observed measures, the variance of the factors then being determined by the known reliabilities. In the typical achievement study the true variance increases over time ($V_{Y_2} > V_{Y_1}$), e.g., because some students will pursue the study of mathematics whereas others will avoid advanced courses. The usual factor and path analysis approach of standardizing all factors (e.g., $V_{Y_1} = V_{Y_2}$) is clearly unsatisfactory for growth studies because it ignores changes in true variance. Even if there were no errors of measurement, standardization of variables is undesirable in growth studies. Psychometricians have usually dealt with models in which one measure of a construct was available, but when several measures with different units are obtained the variance of the construct becomes arbitrary. If the initial status factor is assigned a variance of unity ($V_{Y_1} = 1$) then the assumption $b_1 = b_2$ will identify the variance of the final status factor (Werts & Linn, 1970b) given that b_1^* and b_2^* are identified. Werts, Jöreskog, and Linn (1972) show that if we have, e.g., two congeneric measures of Y_1 (X_{11} and X_{12}) which are repeated at a later time (X_{21} and X_{22} respectively), then it is possible to test whether the assumption $b_1 = b_2$ is compatible with $b_3 = b_4$. In other words the ratio $V_{Y_2} : V_{Y_1}$ identified by the assumption that $b_1 = b_2$ may be different from the ratio of these variances given by the assumption that $b_3 = b_4$ and this will show up as a significant increase in lack of fit of the model to the data when the added assumption $b_3 = b_4$ is imposed on the model. This test indicates whether it is reasonable to believe that both measures have equal units over time.

7. Other Constructs in Statistical Procedures

In this section we propose to demonstrate that statistical procedures frequently imply constructs which many researchers are not aware of. For illustrative purposes consider a quasi-experimental (Campbell & Stanley, 1963) study in which four different procedures for teaching fifth grade mathematics are randomly assigned to four available schools in a district. The mathematics achievement of each student is measured at the beginning and end of fifth grade using parallel forms of a test which provide good coverage of the material taught in the various schools (i.e., the test has face validity). As frequently happens in naturalistic studies it is found that the mean achievement scores at the beginning of the fifth grade differ. To avoid interpretive complications assume perfect validity. Suppose that the mean results for schools are as shown in Fig. 5, i.e., the ordering of the schools remained constant over time but the spread of means increased in proportion to the initial mean. One possible statistical procedure which the data seem to fit is the analysis of variance of repeated measures (Winer, 1962, Chap. 7) which basically consists of subtracting the initial means from the final means and testing to see if these differences are the same from school to school. Since these differences range from 20 units to 5 units for schools #1 and #4 respectively, it is clear that this procedure would conclude that there is a treatment (i.e., school) effect, i.e., school #1 is the most and #4 the least effective. A second statistical procedure which the data fit is the analysis of covariance with initial status controlled (Winer, 1962, Chap. 11). Since the final means are perfectly correlated with the initial means it may be shown that this procedure will indicate no treatment (i.e., school) effect, given the standard analysis of covariance assumptions (Werts & Linn, 1972). In order to understand these seemingly contradictory interpretations, we need to ponder the following hypothetical question: For any given school, what would the final mean be if no treatment had been applied? The analysis of variance in essence assumes that for each school, if no treatment had been given, then the final mean would be the same as the initial mean. In contrast the analysis of covariance assumes that if no treatment were given then the final mean would be completely predictable from the initial mean, i.e., in our illustration the final means are perfectly correlated with initial means. There is no law of nature that either case is necessarily so, which means that neither statistical procedure may be appropriate. Furthermore, our analysis has assumed the appropriateness of a linear addition model, which may not provide a reasonable simulation of the reality being investigated.

A slight variation in the above problem occurs when some measure is being obtained in a time series and at some point a new treatment is imposed. Such a case might be in the math achievement of students who are being followed from grade school into high school. Thistlethwaite and Campbell (1960) have argued that if the past treatment trend continues on the pretreatment trend then no treatment effect may be inferred. In real life, however, students who go to a superior high school have probably gone to superior grade schools and vice versa. If so, then it is quite possible that the effective high school would do well if it could continue the learning progress its students were making before entry. A treatment effect might well be evidenced by a straight trend line from grade school through high school. Again, the unobserved construct is: What would the group mean be if there were no treatment? Without this information no statements about treatment effects are warranted, nor can anybody validly assert that a particular statistical analysis is appropriate,

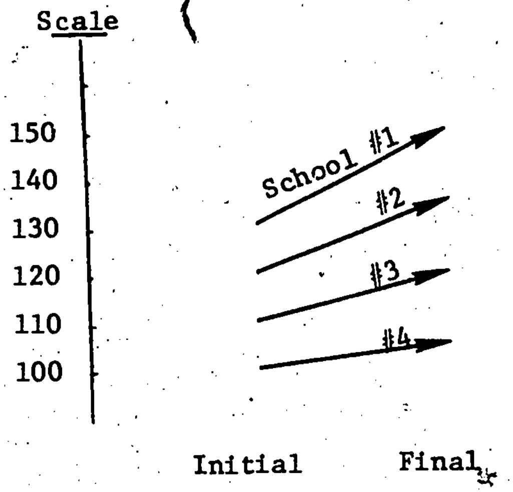


Fig. 5. Mean math scores.

except within the context of a particular model with its associated assumptions.

8. Hypotheses About Changes in Means

The discussion to this point has been devoted to the analysis of the observed variance-covariance matrix. In some problems, however, hypotheses really concern structures (i.e., restrictions) on the means of variables, e.g., if we gave a class some special assistance in vocabulary we would like to observe an increase in the average vocabulary score of the group, i.e., the correlation between initial and final vocabulary scores would not be the relevant statistic to analyze. In such cases the neglect of means (common among path analysts) would lead to uninterpretable results.

Educational researchers interested in growth have encountered the problem of means because of the way that tests are constructed (e.g., Carver, 1970). The procedures used in test development typically strive to maximize the discrimination between individuals, e.g., items that are answered correctly by almost everyone at the end of a course tend to be omitted since these serve to show similarities among individuals. Yet it may be precisely these items that show the general progress of the class during the course. The item analysis procedures thus prevent measurement of true change in means over time. Consider the extreme case in which the students have no familiarity with the subject matter being taught, which would mean that an initial test of their knowledge in this subject would yield a zero score for the whole class. A parallel test given at the end of the course would show varying degrees of knowledge attained, i.e., a positive mean and variance. The initial test scores would be expected to have a zero (meaningless) correlation with the final scores and the final mean would represent the average level of course effectiveness. If initially students had little or no familiarity with the subject matter then the reliability of the initial test might be quite low and yet this measure might be appropriate for measuring changes in student knowledge during the course. Obviously path coefficients would be irrelevant to the issue.

As noted above, parallel tests are assumed to have the same underlying mean. Thus, underlying the various observed test means, there is assumed to be a common true score mean. If the means do not differ significantly, then the best estimate of the true mean is the grand mean of the observed tests. Notice that if the grand mean is used as the best estimate of the common test mean, then this will affect our estimates of variances and covariances since these are measures of deviation from the grand mean. This mutual interdependence is recognized in Jöreskog's (1970) general model, which allows for simultaneous estimation and hypothesis testing given restrictions on both means and the variance-covariance matrix. We may, for example, wish to test the hypothesis that the true score means over time increase linearly (or exponentially).

9. General Considerations

It is relatively easy to find a linear structural model which fits the data quite closely, e.g., factor analysts may keep adding factors until a good fit is obtained. With a modicum of thought it is also relatively easy to obtain a model which is consistent with our theory, when this model is just identified (i.e., there is a unique solution for each parameter), because the matrix estimated from the model $(\hat{\Sigma})$ will in general equal the observed matrix (S). Given overidentification, it is possible that the model may be rejected because of poor fit to the data. In such cases it is usually possible to find a less restrictive model which will fit the data better, but this model may not be substantively plausible. It is extremely difficult to demonstrate that (a) a model approximately simulates reality, (b) it provides better simulation than another model, (c) the constructs defined by the model have greater explanatory power than the observed variables from which they are derived, and (d) these constructs are in any sense useful in promoting better research. In most cases it seems reasonable to suppose that several plausible models may be found, all of which are consistent with the observed data. It would then be necessary to deduce what data would need to be collected to discriminate among these models.

Some of the concepts discussed in previous sections suggest some cautions in interpreting observed variance-covariance matrices. Granting the validity of using correlations at all (see Tukey, 1954, for a discussion of this question), it should be clear from the section on the multitrait-multimethod procedure that the probable existence of errors of measurement and multiple indicators of underlying variables will necessarily make any interpretation a chancy affair. Furthermore, even if the unattenuated correlations among the relevant constructs were known, correlations are by no means self-interpreting in a causal sense (Blalock, 1964). Thus an observed correlation may be completely spurious due to the presence of a common antecedent variable (which must be controlled). While most psychologists use the concept of spuriousness, the notion of controlling a variable in a chain of causes to see if this variable explains the observed association (Blalock, 1964) is almost unknown at present. It should not be inferred, however, that a causal analysis of the correlations is appropriate to every problem (Bailey, 1970).

Most applications of factor analysis, path analysis, and test theory can probably be described as exploratory or speculative in the sense that the analysis was performed because the researcher was familiar with that technique rather than because it could be demonstrated that his approach provided a better simulation of the process under study. We are thus in the unenviable position of discussing statistical techniques without knowing when they should be used. The value of these techniques has yet to be demonstrated in most of the social sciences with the possible exception of economics.

References

Bailey, K. D. Evaluating axiomatic theories. In E. F. Borgatta & G. W. Bohrnstedt (Eds.), Sociological methodology, 1970. San Francisco: Jossey-Bass, 1970.

Blalock, H. M., Jr. Causal inferences in nonexperimental research. Chapel Hill: University of North Carolina Press, 1964.

Blalock, H. M., Jr. The identification problem and theory building. American Sociological Review, 1966, 31, 52-61.

Blalock, H. M., Jr. Causal inferences, closed populations, and measures of association. American Political Science Review, 1967, 61, 130-136.

Blalock, H. M., Jr. Multiple indicators and the causal approach to measurement error. The American Journal of Sociology, 1969, 75, 264-272.

Blalock, H. M., Wells, C. S., & Carter, L. F. Statistical estimation with random measurement error. In E. F. Borgatta & G. W. Bohrnstedt (Eds.), Sociological methodology, 1970. San Francisco: Jossey-Bass, 1970.

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

Boruch, R. F., Larkin, J. D., Wolins, L., & McKinney, A. C. Alternative methods of analysis: multitrait-multimethod data. Educational and Psychological Measurement, 1970, 30, 833-854.

Boudon, R. A method of linear causal analysis: dependence analysis. American Sociological Review, 1965, 30, 365-374.

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

Campbell, D. T., & Stanley, J. S. Experimental and quasi-experimental designs for research in teaching. In N. L. Gage (Ed.), Handbook of research on teaching. Chicago: Rand McNally, 1963.

Carver, R. P. A model for using the final examination as a measure of the amount learned in classroom learning. Journal of Educational Measurement, 1969, 6, 59-68.

Carver, R. P. Special problems in measuring change with psychometric devices. In Evaluation research: Strategies and methods. Pittsburgh: American Institute of Research, 1970. Pp. 48-66.

Costner, H. L. Theory, deduction, and rules of correspondence. The American Journal of Sociology, 1969, 75, 245-263.

Cronbach, L. J., & Furby, L. How should we measure "change"--or should we? Psychological Bulletin, 1970, 74, 68-80.

Duncan, O. D. Path analysis: Sociological examples. The American Journal of Sociology, 1966, 72, 1-16.

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

Goldberger, A. S. Econometrics and psychometrics: A survey of communalities. Psychometrika, 1970, 36, 83-107.

Guilford, J. P. Psychometric methods. New York: McGraw-Hill, 1954.

Guttman, L. Reliability formulas that do not assume experimental independency. Psychometrika, 1953, 18, 225-239.

Harman, H. H. Modern factor analysis. Chicago: University of Chicago Press, 1967.

Harris, C. W. Problems in measuring change. Madison: University of Wisconsin Press, 1963.

Hauser, R. M., & Goldberger, A. S. The treatment of unobservable variables in path analysis. In E. F. Borgatta & G. W. Bohrnstedt (Eds.), Sociological methodology, 1970. San Francisco: Jossey-Bass, 1970.

Isaac, P. D. Linear regression, structural relations, and measurement error. Psychological Bulletin, 1970, 74, 213-218.

Johnston, J. Econometric methods. New York: McGraw-Hill, 1963.

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

Jöreskog, K. G. Statistical analysis of sets of congeneric tests. Psychometrika, 1971, 36, 109-134.

Kendall, M. G., & Stuart, A. The advanced theory of statistics. Vol. II. Inference and relationship. London: Griffin, 1961.

Lazarsfeld, P. F. The logical and mathematical foundation of latent structure analysis. In S. Stauffer et al. (Eds.), Studies on social psychology in World War II. Vol. 4. Measurement and prediction. Princeton, N. J.: Princeton University Press, 1950.

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

Stanley, J. C. Reliability. In R. L. Thorndike (Ed.), Educational measurement. Washington, D. C.: American Council on Education, 1971. Pp. 356-442.

Theil, H. Specification errors and the estimation of economic relationships. Review International Statistics Institute, 1957, 25, 41-51.

Thistlethwaite, D. L., & Campbell, D. T. Regression-discontinuity analysis: An alternative to the ex post facto experiment. Journal of Educational Psychology, 1960, 51, 309-317.

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

Tukey, J. W. Causation, regression, and path analysis. In O. Kempthorne et al. (Eds.), Statistics and mathematics in biology. Ames, Iowa: Iowa State College Press, 1954.

Werts, C. E., Jöreskog, K. G., & Linn, R. L. A multitrait-multimethod model for studying growth. Educational and Psychological Measurement, 1972 (in press).

Werts, C. E., & Linn, R. L. Cautions in applying various procedures for determining the reliability and validity of multiple-item scales. American Sociological Review, 1970, 35, 757-759. (a)

Werts, C. E., & Linn, R. L. Path analysis: Psychological examples. Psychological Bulletin, 1970, 74, 193-212. (b)

Werts, C. E., & Linn, R. L. Corrections for attenuation. Educational and Psychological Measurement, 1971, 32, 117-127.

Werts, C. E., & Linn, R. L. Problems with inferring treatment effects from repeated measures. Educational and Psychological Measurement, 1972, in press.

Winer, B. J. Statistical principles in experimental design. New York: McGraw-Hill, 1962.

Wright, S. The method of path coefficients. *Annals of Mathematical Statistics*, 1934, *5*, 161-215.

IV. A Synthesis of Psychometric Literature: A Multitrait-multimethod Model for Studying Growth

Werts and Linn (1970a) have suggested that a multitrait-multimethod approach (Campbell & Fiske, 1959) might be used for studying growth. The purpose of this paper is to detail such a model and to outline implications for the study of growth. The major focus of our exposition will be the logic of this model rather than the estimation of parameters or testing the fit of the model to data. A comprehensive discussion of appropriate estimation and fit-testing procedures may be found in Jöreskog (1970a), whose general model for the analysis of covariance structures subsumes the models used in this paper.

The Model

The multitrait-multimethod approach may be treated as a problem in confirmatory factor analysis (Jöreskog, 1970a, 1971). For illustrative purposes we will consider the example of three traits and three methods since this is the minimum number of traits and methods required to produce unique (defined in Jöreskog, 1969, pp. 185-186) parameter estimates, given the assumption that each observed measure loads on only one trait and one method factor and all factors are oblique. The general factor analytic model is:

$$\underset{\sim}{y} = \underset{\sim}{\mu} + \underset{\sim}{\Lambda}\underset{\sim}{T} + \underset{\sim}{e} \qquad (1)$$

where $\underset{\sim}{y}$ is the vector of observed scores,

$\underset{\sim}{\mu}$ is the mean vector of $\underset{\sim}{y}$,

$\underset{\sim}{\Lambda}$ is a matrix of factor loadings,

$\underset{\sim}{T}$ is a vector of common factor scores, and

$\underset{\sim}{e}$ is a vector of unique factor scores corresponding to specific factors and/or errors of measurement.

For our example:

$$\underset{\sim}{y}' = (y_{11}, y_{21}, y_{31}, y_{12}, y_{22}, y_{32}, y_{13}, y_{23}, y_{33}) \quad , \qquad (1a)$$

where in y_{ij} , i = method and j = trait,

$$\underset{\sim}{T}' = (T_1, T_2, T_3, M_1, M_2, M_3) \quad , \qquad (1b)$$

where T_j = the j -th trait factor,

M_i = the i -th method factor,

$$\underset{\sim}{\Lambda} = \begin{bmatrix} A_{11} & 0 & 0 & B_{11} & 0 & 0 \\ A_{21} & 0 & 0 & 0 & B_{21} & 0 \\ A_{31} & 0 & 0 & 0 & 0 & B_{31} \\ 0 & A_{12} & 0 & B_{12} & 0 & 0 \\ 0 & A_{22} & 0 & 0 & B_{22} & 0 \\ 0 & A_{32} & 0 & 0 & 0 & B_{32} \\ 0 & 0 & A_{13} & B_{13} & 0 & 0 \\ 0 & 0 & A_{23} & 0 & B_{23} & 0 \\ 0 & 0 & A_{33} & 0 & 0 & B_{33} \end{bmatrix} \qquad (1c)$$

where A_{ij} are loadings on trait factors and

B_{ij} are loadings on method factors.

The expected variance-covariance matrix $\underset{\sim}{\Sigma}$ of $\underset{\sim}{y}$ is then given by

$$\underset{\sim}{\Sigma} = \underset{\sim}{\Lambda}\underset{\sim}{\Phi}\underset{\sim}{\Lambda}' + \theta^2 \qquad (2)$$

where θ^2 is a diagonal matrix whose elements are the variances of $\underset{\sim}{e}$. Since all factors are oblique, in our example:

$$\Phi = \begin{bmatrix} V_{T_1} & & \text{Symmetric} & & & \\ C_{T_1T_2} & V_{T_2} & & & & \\ C_{T_1T_3} & C_{T_2T_3} & V_{T_3} & & & \\ C_{T_1M_1} & C_{T_2M_1} & C_{T_3M_1} & V_{M_1} & & \\ C_{T_1M_2} & C_{T_2M_2} & C_{T_3M_2} & C_{M_1M_2} & V_{M_2} & \\ C_{T_1M_3} & C_{T_2M_3} & C_{T_3M_3} & C_{M_1M_3} & C_{M_2M_3} & V_{M_3} \end{bmatrix} \qquad (2a)$$

where the C's are covariances and the V's are variances.

Following Jöreskog (1970a), parameters will be labelled as one of three kinds: (1) fixed parameters that have been assigned given values; (2) constrained parameters that are unknown but equal to one or more other parameters; and (3) free parameters that are unknown and not constrained to be equal to any other parameter. The term "identifiable" will be used in the sense defined by Fisher (1966, p. 25): "we shall speak of that equation as identifiable (or identified) if there exists some combination of prior and posterior information which will enable us to distinguish its parameters from those of any other equation in the same form." For the models studied in this paper, the term "identifiable" is synonymous with the factor analyst's term "unique solution," i.e., a solution is "unique" if all linear transformations of the factors that leave the fixed parameters unchanged also leave the free parameters unchanged. As Jöreskog (1970b) notes: "Before an attempt is made to estimate a model of this kind, the identification problem must be examined." The number of overidentifying restrictions on the model is frequently of interest, for example, after standardizing factor variances (i.e., $V_{T_j} = V_{M_i} = 1$) the three method by three trait model has three overidentifying restrictions, i.e., Σ has 45 distinct variances and covariances as compared to 42 free parameters to be estimated (18 factor loadings, 15 factor covariances in Φ, and nine residual variances in θ). The number of overidentifying restrictions are the degrees of freedom (df) for the test statistic in Jöreskog's general model (1970a, p. 241, sec. 1.4). The "path analysis" approach used by Werts and Linn (1970a) can be very useful in exploring the identification question in overidentified models. However, as noted by Hauser and Goldberger (1970) the "path analysis" literature does not adequately deal with the estimation problem in overidentified models, in part because the sample-population distinction is blurred.

The multitrait-multimethod approach considered above does not consider any functional relationships among the trait factors, i.e., the approach deals only with errors of measurement. In the study of growth, these trait factors correspond to initial status, final status, and the determinants of growth and a structural model showing the relationship among these variables must be specified. Substantive inferences about growth are based on estimates of the parameters of the structural model.

Suppose that the structural model for growth took the form:

$$T_3 = D_1T_1 + D_2T_2 + \xi \qquad (3)$$

where T_3 is the final status, T_2 is the initial status, and T_1 is a determinant of growth; all other influences on growth (represented by ξ) being independent of T_1 and T_2. In this model the initial status T_2 may influence the rate of growth. The parameters of equation (3) are just identifiable in terms of the elements of Φ, i.e., the number of restrictions on the overall model is not changed. Assuming that T_3 and T_2 are measurements on the same dimension as implied by the terms "initial" and "final" status, growth (Δ) is equal to $T_3 - T_2$. Werts and Linn (1970b) have shown that the regression weights for T_1 and T_2 are:

$$D_1 = D_{\Delta T_1 . T_2} \quad , \tag{4}$$

and

$$D_2 = 1 + D_{\Delta T_2 . T_1} \tag{5}$$

where $D_{\Delta T_1 . T_2}$ is the regression weight of Δ on T_1 with T_2 controlled and $D_{\Delta T_2 . T_1}$ is the regression weight of Δ on T_2 with T_1 controlled. In other words D_1 represents the direct influence of T_1 on growth and D_2 represents the direct influence of initial status on growth plus unity (which represents that part of T_3 which is initial status). Since $T_3 = \Delta + T_2$, substituting equations (4) and (5) into (3) yields:

$$\Delta = D_{\Delta T_1 . T_2} T_1 + D_{\Delta T_2 . T_1} T_2 + \xi \quad . \tag{6}$$

In terms of T_1 , T_2 and ξ , equations (1b), (1c), and (2a) become:

$$\underset{\sim}{T}^* = (T_1, T_2, \xi, M_1, M_2, M_3) \quad , \tag{7a}$$

$$\underset{\sim}{\Lambda}^* = \begin{bmatrix} A_{11} & 0 & 0 & B_{11} & 0 & 0 \\ A_{21} & 0 & 0 & 0 & B_{21} & 0 \\ A_{31} & 0 & 0 & 0 & 0 & B_{31} \\ 0 & A_{12} & 0 & B_{12} & 0 & 0 \\ 0 & A_{22} & 0 & 0 & B_{22} & 0 \\ 0 & A_{32} & 0 & 0 & 0 & B_{32} \\ A_{13}D_1 & A_{13}D_2 & A_{13} & B_{13} & 0 & 0 \\ A_{23}D_1 & A_{23}D_2 & A_{23} & 0 & B_{23} & 0 \\ A_{33}D_1 & A_{33}D_2 & A_{33} & 0 & 0 & B_{33} \end{bmatrix} \quad , \tag{7b}$$

and

$$\Phi^* = \begin{bmatrix} 1 & & & \text{Symmetric} & & \\ C_{T_1T_2} & 1 & & & & \\ 0 & 0 & V_\xi & & & \\ C_{T_1M_1} & C_{T_2M_1} & C_{\xi M_1} & 1 & & \\ C_{T_1M_2} & C_{T_2M_2} & C_{\xi M_2} & C_{M_1M_2} & 1 & \\ C_{T_1M_3} & C_{T_2M_3} & C_{\xi M_3} & C_{M_1M_3} & C_{M_2M_3} & 1 \end{bmatrix} , \qquad (7c)$$

respectively. If the analyst wished to scale a factor by the unit of a particular measure this may be accomplished by setting the A_{ij} slope for the measure equal to unity (in which case the variance of the corresponding factor should not be standardized but left free to be estimated by the program). The assumption that T_2 and T_3 are measures on the same dimension is equivalent to setting the same method regression weights equal, i.e., in our example $A_{12} = A_{13}$, $A_{22} = A_{23}$ and $A_{32} = A_{33}$. As detailed by Werts and Linn (1970a) the effect of these restrictions is that the ratio of the variance of T_3 to T_2 is fixed. For estimation purposes it is convenient to standardize all factors except T_3 whose variance is fixed in relation to T_2. The model defined by equations (7a), (7b), and (7c) is no longer a simple factor analysis model, but may be estimated using Jöreskog's (1970a) general model for the analysis of covariance structures. For this purpose $\underset{\sim}{\Lambda}^*$ may be rewritten as the product of two matrices:

$$\underset{\sim}{\Lambda}^* = \underset{\sim}{B}\underset{\sim}{\Lambda}^{**}$$

where

$$\underset{\sim}{B} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & A_{13} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & A_{23} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & A_{33} \end{bmatrix}$$

and

$$\Lambda^{**} = \begin{bmatrix} A_{11} & 0 & 0 & B_{11} & 0 & 0 \\ A_{21} & 0 & 0 & 0 & B_{21} & 0 \\ A_{31} & 0 & 0 & 0 & 0 & B_{31} \\ 0 & A_{12} & 0 & B_{12} & 0 & 0 \\ 0 & A_{22} & 0 & 0 & B_{22} & 0 \\ 0 & A_{32} & 0 & 0 & 0 & B_{32} \\ D_1 & D_2 & 1 & \chi_{13} & 0 & 0 \\ D_1 & D_2 & 1 & 0 & \chi_{23} & 0 \\ D_1 & D_2 & 1 & 0 & 0 & \chi_{33} \end{bmatrix}$$

and $\chi_{13} = B_{13}/A_{13}$, $\chi_{23} = B_{23}/A_{23}$, $\chi_{33} = B_{33}/A_{33}$. By substitution:

$$\underset{\sim}{\Sigma} = \underset{\sim}{B}\underset{\sim}{\Lambda}^{**}\underset{\sim}{\Phi}^{*}\underset{\sim}{\Lambda}^{**\prime}\underset{\sim}{B}' + \underset{\sim}{\theta}^{2} \quad ,$$

which is a special case of Jöreskog's (1970a) general model.

In using the computer program (Jöreskog, Gruvaeus, & van Thillo, 1970) the parameters A_{12} , A_{22} , A_{32} in Λ^{**} should be constrained to be equal to A_{13} , A_{23} , and A_{33} respectively in $\underset{\sim}{B}$. The resulting model has 45 distinct variances and covariances in $\underset{\sim}{\Sigma}$ and 40 free and constrained parameters (17 in $\underset{\sim}{\Lambda}^{**}$, 14 in $\underset{\sim}{\Phi}^{*}$, 9 in $\underset{\sim}{\theta}$, none in B because of equality restraints), which means that the model has five overidentifying restrictions (df). The advantage of casting the analysis in terms of Jöreskog's general model is that, given the assumption that the observed variables are distributed normally, various hypotheses about the model may be tested in large samples. In particular, we may wonder if trait factors are uncorrelated with methods factors and methods factors with each other as assumed by Cronbach and Furby (1970) and Werts and Linn (1970a) in their analysis of growth. To make this test, the analysis would be run with the model of (1a), (1b), and (1c), and (2a) with $V_{T_1} = V_{T_2} = V_{T_3} = V_{M_1} = V_{M_2} = V_{M_3} = 1$ and then the analysis would be made with $C_{T_1M_1} = C_{T_1M_2} = C_{T_1M_3} = C_{T_2M_1} = C_{T_2M_2} = C_{T_2M_3} = C_{T_3M_1} = C_{T_3M_2} = C_{T_3M_3} = C_{M_1M_2} = C_{M_1M_3} = C_{M_2M_3} = 0$. For our example, the initial analysis would yield a chi-square with three df for testing the fit of the model to the data. The second analysis would yield a chi-square with 15 df since 12 additional restrictions have been made. The increase in chi-square with 12 df is a test of the tenability of the additional restrictions. Starting with the same initial model, the tenability of assuming that $A_{12} = A_{13}$, $A_{22} = A_{23}$, and $A_{32} = A_{33}$ may be tested (dropping the $V_{T_3} = 1$ assumption) using the increase in chi-square with 2 df. Likewise starting with these assumptions (i.e., equations (7a), (7b), and (7c), and df = 5) hypotheses about growth can be tested, e.g., D_1 can be set equal to zero and the resulting change in χ^2 (df = 1) is a test of whether T_2 directly influences growth. To test whether initial status directly influences growth (i.e., whether $D_{\Delta T_2 . T_1} = 0$), D_2 would be set equal to unity (see equation (4)), the increase in χ^2 (df = 1) testing this hypothesis. The fit of the observed variance-covariance matrix S to the estimated elements of Σ may be used to form some judgment as to changes in fit resulting from additional restrictions, especially when the χ^2 test is inappropriate because the assumption of multivariate normality is not reasonable.

As originally conceived by Campbell and Fiske (1959) the multitrait-multimethod approach required each trait to be measured with each method, as in the example analyzed above. The linear structural model approach proposed herein requires that model parameters be identifiable, a question which is unrelated to whether each trait is measured with each method. In order to fix the ratio of the variance of the final status to the initial status factor, only one pair of initial and final measures with the same units of measurement are required, i.e., the three sets of initial-final measures in our example serve to overidentify this variance ratio. The identification problem would be greatly simplified if one of these same method sets were replaced with different method measures, even though the resulting matrix would no longer be in the form required by Campbell and Fiske. Campbell and Fiske's argument that different method measures of a trait are required to improve convergent validity appears fundamentally sound and is a basic premise in our analysis. We have abandoned the particular type of analysis used by Campbell and Fiske because it fails to specify the underlying structure being postulated, and does not allow for nonsymmetrical method-by-trait combinations.

Relationship to Classical Test Theory

The multitrait-multimethod formulation can be shown to include various procedures derived from classical test theory as special cases, e.g., the commonly used formulas for reliability of differences, correlation of true initial status with true gain, and the correlation of true scores over time can be derived from the multitrait-multimethod model by imposing specifiable restrictions. To illustrate this point we shall examine the case of two parallel measures (y_{12}, y_{22}) given initially and two finally (y_{13}, y_{23}). First let us consider the analysis given the traditional assumptions that all errors of measurement are independent of each other and of the true scores. In our formulation this is equivalent to asserting that there are no methods factors. Without further assumptions the model may be represented in terms of equation (1) as

$$\underset{\sim}{y} = (y_{12}, y_{22}, y_{13}, y_{23}) \quad , \qquad (8a)$$

$$\underset{\sim}{T} = (T_2, T_3) \quad , \qquad (8b)$$

$$\underset{\sim}{\Lambda} = \begin{bmatrix} A_{12} & 0 \\ A_{22} & 0 \\ 0 & A_{13} \\ 0 & A_{23} \end{bmatrix} \quad , \qquad (8c)$$

$$\underset{\sim}{\Phi} = \begin{bmatrix} V_{T_2} & \\ C_{T_2 T_3} & V_{T_3} \end{bmatrix} \quad , \qquad (8d)$$

and

$$\underset{\sim}{\theta}^2 = \begin{bmatrix} V_{e_{12}} & & & \\ 0 & V_{e_{22}} & & \\ 0 & 0 & V_{e_{13}} & \\ 0 & 0 & 0 & V_{e_{23}} \end{bmatrix} \quad . \qquad (8e)$$

Assuming that initial and final status are on the same scale, "parallel" test assumptions are equivalent to (Jöreskog, 1971) fixing $A_{12} = A_{22} = A_{13} = A_{23} = 1$ and constraining $V_{e_{12}} = V_{e_{22}}$ and $V_{e_{13}} = V_{e_{23}}$.

All parameters are identifiable and df = 5 . Identification still occurs without the error variance assumptions (df = 3) , i.e., in true score lexicon, "essentially tau-equivalent" measures (Lord & Novick, 1968, pp. 47-50) would suffice. If we choose to use nonparallel or "congeneric" (Jöreskog, 1971) measures, one pair of measures over time being on the same scale (e.g., $A_{12} = A_{13}$) , V_{T_2} could be arbitrarily standardized (= 1), yielding an identifiable model with df = 1 . In all these cases, growth statistics may be obtained from the parameter estimates or the model can be transformed to obtain growth statistics directly. Inserting $T_3 = T_2 + \Delta$ then:

$$T^* = (T_2 \, , \, \Delta) \quad , \qquad (9a)$$

$$\underset{\sim}{\Lambda}^* = \begin{bmatrix} A_{12} & 0 \\ A_{22} & 0 \\ A_{13} & A_{13} \\ A_{23} & A_{23} \end{bmatrix} , \qquad (9b)$$

where $A_{12} = A_{13}$ by assumption, and

$$\underset{\sim}{\phi}^* = \begin{bmatrix} V_{T_2} & \\ C_{T_2\Delta} & V_\Delta \end{bmatrix} , \qquad (9c)$$

where $V_{T_2} = 1$ for convenience.

Relevant growth statistics are:

$\hat{\rho}_{T_2\Delta}$ = correlation of initial status with gain = $\hat{C}_{T_2\Delta} \div \sqrt{\hat{V}_\Delta \hat{V}_{T_2}}$, (10a)

$$\hat{D}_{\Delta T_2} = \hat{C}_{T_2\Delta} \div \hat{V}_{T_2} , \qquad (10b)$$

$$\hat{V}_{T_3} = \hat{V}_{T_2} + \hat{V}_\Delta + 2\hat{C}_{T_2\Delta} , \text{ and} \qquad (10c)$$

$$\hat{D}_{T_3T_2} = 1 + \hat{D}_{\Delta T_2} . \qquad (10d)$$

Similarly if parameter estimates were derived from the original model of equations (8a), (8b), (8c), (8d), and (8e), growth statistics can be obtained by

$$\hat{D}_{T_3T_2} = \hat{C}_{T_2T_3} \div \hat{V}_{T_2} , \qquad (11a)$$

$$\hat{D}_{\Delta T_2} = \hat{D}_{T_3T_2} - 1 , \qquad (11b)$$

$$\hat{V}_\Delta = \hat{V}_{T_2} + \hat{V}_{T_3} - 2\hat{C}_{T_2T_3} , \qquad (11c)$$

$$\hat{C}_{T_2\Delta} = \hat{D}_{\Delta T_2}\hat{V}_\Delta , \text{ and} \qquad (11d)$$

$$\hat{\rho}_{T_2\Delta} = \hat{D}_{\Delta T_2}\sqrt{\hat{V}_{T_2} \div \hat{V}_\Delta} . \qquad (11e)$$

Following Jöreskog (1971) the parallel test assumption can be tested (given multivariate normality) by comparing the chi-square for the "essentially tau-equivalent" model to that for the "parallel" test model; the difference in chi-square with df = 2 is a test of assumptions that $V_{e_{12}} = V_{e_{22}}$ and $V_{e_{13}} = V_{e_{23}}$. Similarly the increase in chi-square from the "congeneric" model to the "essentially tau-equivalent" model (df = 2) is a test of the assumptions that $A_{12} = A_{22}$ and $A_{13} = A_{23}$. If the parallel test assumptions are accepted then the population reliability at the initial time may be estimated by $\hat{V}_{T_2} \div (\hat{V}_{T_2} + \hat{V}_{e_{12}})$ and reliability at the final time by $\hat{V}_{T_3} \div (\hat{V}_{T_3} + \hat{V}_{e_{13}})$. The reliability for each test is the square of the corresponding standardized factor loading in the case of "essentially tau-equivalent" or "congeneric" measures. Another statistic of interest in the traditional psychometric literature is the reliability of differences (ρ_Δ) which is defined as the true variance of the differences divided by the variance of the observed differences. In the parallel case the estimated population error variances can be used to obtain $\hat{\rho}_\Delta$ directly:

$$\hat{\rho}_\Delta = \frac{\hat{V}_\Delta}{\hat{V}_\Delta + \hat{V}_{e_{12}} + \hat{V}_{e_{13}}} . \qquad (12a)$$

With "essentially tau-equivalent" assumptions no statement is made about equality of error variances so that four reliabilities may be estimated:

$$\hat{\rho}'_\Delta = \frac{\hat{V}_\Delta}{\hat{V}_\Delta + \hat{V}_{e_{12}} + \hat{V}_{e_{13}}} , \qquad (12b)$$

$$\hat{\rho}''_\Delta = \frac{\hat{V}_\Delta}{\hat{V}_\Delta + \hat{V}_{e_{12}} + \hat{V}_{e_{23}}} , \qquad (12c)$$

$$\hat{\rho}'''_\Delta = \frac{\hat{V}_\Delta}{\hat{V}_\Delta + \hat{V}_{e_{22}} + \hat{V}_{e_{13}}} , \qquad (12d)$$

$$\hat{\rho}''''_\Delta = \frac{\hat{V}_\Delta}{\hat{V}_\Delta + \hat{V}_{e_{22}} + \hat{V}_{e_{23}}} . \qquad (12e)$$

Formulas (12a), (12b), (12c), (12d), and (12e) are based on the assumption that the true scores have the same units as the observed scores, which is not true in the case of congeneric measures. Since the regression of observed on true differences is equal to the regression of observed on true scores (Werts & Linn, 1970a, equation (25)) it is only necessary to standardize this weight with the appropriate variances to obtain the reliability of differences for all cases, e.g., in the congeneric case if $A_{12} = A_{13}$ then

$$\hat{\rho}_{\Delta} = \hat{A}_{12}^{2} \frac{\hat{V}_{\Delta}}{\hat{V}_{y_{12}} + \hat{V}_{y_{13}} - 2\hat{C}(y_{12}, y_{13})} \qquad (12f)$$

where $\hat{V}_{y_{12}}$, $\hat{V}_{y_{13}}$ and $\hat{C}(y_{12}, y_{13})$ are the estimated elements in $\hat{\Sigma}$. This formula uses estimated elements in $\hat{\Sigma}$ which are provided in the computer output for Jöreskog's program (Jöreskog, Gruvaeus, & van Thillo, 1970). The program computes the elements in $\hat{\Sigma}$ from the estimates for the underlying parameters, e.g., $\hat{C}(y_{12}, y_{13}) = \hat{A}_{12}\hat{A}_{13}\hat{C}_{T_2T_3}$.

This model (all measurement errors independent) may be used to clarify traditional procedures for obtaining growth statistics. For example, consider the case in which one initial and one final test is given. A common procedure is to obtain split half reliabilities at each time and use these to correct for attenuation. If y_{12} and y_{22} are the initial split halves and y_{13} and y_{23} the final split halves, this case corresponds exactly to the parallel measure case analyzed above. The difference from the traditional procedure is that the complete variance-covariance matrix for the split halves is computed and used in the analysis. As shown above, the "parallel" and "essentially tau-equivalent" assumptions can be tested against the congeneric model and the congeneric model is overidentified. From this perspective the traditional procedure neglects useful information about correlations among split halves and thereby loses the possibility of rejecting the model because of poor fit to the data and of analyzing the data making only congeneric test assumptions. To understand the connection with the traditional formula it is of interest to standardize $\hat{\Sigma}$ into a correlation matrix (correlations generated by the model are indicated by symbol ρ) and to show the relationships to standardized model parameters (denoted by asterisk):

$$\rho(y_{12}, y_{13}) = \hat{A}^{*}_{12}\hat{\rho}_{T_2T_3}\hat{A}^{*}_{13} \qquad (13a)$$

$$\rho(y_{12}, y_{23}) = \hat{A}^{*}_{12}\hat{\rho}_{T_2T_3}\hat{A}^{*}_{23} \qquad (13b)$$

$$\rho(y_{22}, y_{13}) = \hat{A}^{*}_{22}\hat{\rho}_{T_2T_3}\hat{A}^{*}_{13} \qquad (13c)$$

$$\rho(y_{22}, y_{23}) = \hat{A}^{*}_{22}\hat{\rho}_{T_2T_3}\hat{A}^{*}_{23} \qquad (13d)$$

$$\rho(y_{12}, y_{22}) = \hat{A}^*_{12}\hat{A}^*_{22} \qquad (13e)$$

$$\rho(y_{13}, y_{23}) = \hat{A}^*_{13}\hat{A}^*_{23} \quad . \qquad (13f)$$

If parallel test assumptions are valid then $\hat{A}^*_{12} = \hat{A}^*_{22}$ and $\hat{A}^*_{13} = \hat{A}^*_{23}$, in which case equations (13a), (13b), (13c), and (13d) are identical and should be recognized as the traditional correction for attenuation, except that the correlations are drawn from $\hat{\Sigma}$ rather than from the observed correlation matrix $\underset{\sim}{S}$. Equations (13e) and (13f), under parallel test assumptions, are simply the assumption that the reliability defined as the squared correlation (i.e., A^*_{12} or A^*_{13}) of the observed with the true score is equal to the correlation between two parallel tests, but again the correlations are drawn from $\hat{\Sigma}$ not from $\underset{\sim}{S}$. What these equations show is that it is not necessary for the reliabilities of the split halves to be equal in order to identify the unattenuated correlation $\hat{\rho}_{T_2T_3}$ given uncorrelated errors. If the estimates of the elements in $\hat{\Sigma}$ for the parallel case are examined it will be found that because of the structural specifications: $\hat{V}_{y_{12}} = \hat{V}_{y_{22}}$, $\hat{V}_{y_{13}} = \hat{V}_{y_{23}}$, $\hat{C}(y_{12}, y_{13}) = \hat{C}(y_{13}, y_{23}) = \hat{C}(y_{22}, y_{13}) = \hat{C}(y_{22}, y_{23})$, $\hat{C}(y_{12}, y_{22}) = \hat{V}_{T_2}$, $\hat{C}(y_{13}, y_{23}) = \hat{V}_{T_3}$ and $\hat{C}(y_{12}, y_{13}) = \hat{C}(y_{22}, y_{23}) = \hat{C}_{T_2T_3}$. Translating the equation for the reliability of differences into the elements of $\hat{\Sigma}$:

$$\hat{\rho}_{\Delta} = \frac{\hat{C}(y_{12}, y_{22}) + \hat{C}(y_{13}, y_{23}) - 2\hat{C}(y_{12}, y_{13})}{\hat{V}_{y_{12}} + \hat{V}_{y_{13}} - 2\hat{C}(y_{12}, y_{13})} \qquad (14a)$$

or

$$\hat{\rho}_{\Delta} = \frac{\hat{V}_{y_{12}}\hat{\rho}(y_{12}, y_{22}) + \hat{V}_{y_{13}}\hat{\rho}(y_{13}, y_{23}) - 2\hat{\rho}(y_{12}, y_{13})\sqrt{\hat{V}_{y_{12}}\hat{V}_{y_{13}}}}{\hat{V}_{y_{12}} + \hat{V}_{y_{13}} - 2\hat{\rho}(y_{12}, y_{13})\sqrt{\hat{V}_{y_{12}}\hat{V}_{y_{13}}}} \; . \qquad (14b)$$

Equation (14b) should be recognized as the traditional formula for the reliability of differences, noting however that the estimates are drawn from $\hat{\Sigma}$, not from the observed matrix $\underset{\sim}{S}$. The essentially tau-equivalent case differs from the parallel case in that the corresponding variances in $\hat{\Sigma}$ are not required to be equal, however the covariances between independent measures of different traits are still equal to the covariances between the corresponding traits factors. This means that formula (14a) could be used for any pair of tau-equivalent tests over time. For congeneric measures the formula involves the pairs of measures which have the same units over time, e.g., if $A_{12} = A_{13}$ then equation (12f) may be translated into

$$\hat{\rho}_{\Delta} = \frac{\hat{V}_{y_{12}}(\hat{A}^*_{12})^2 + \hat{V}_{y_{13}}(\hat{A}^*_{13})^2 - 2\hat{A}^*_{12}\hat{A}^*_{13}\hat{\rho}_{T_2T_3}\sqrt{\hat{V}_{y_{12}}\hat{V}_{y_{13}}}}{\hat{V}_{y_{12}} + \hat{V}_{y_{13}} - 2\hat{\rho}(y_{12},y_{13})\sqrt{\hat{V}_{y_{12}}\hat{V}_{y_{13}}}} \quad . \quad (14c)$$

Equation (14c) is the reliability of differences formula given by Werts and Linn (1970a, equation (26)) for the case of correlated errors over time for the pair of measurements on the same scale, i.e., the Werts and Linn formula is also appropriate to the independent error case when applied to the elements of $\hat{\underset{\sim}{\Sigma}}$ rather than $\underset{\sim}{S}$. If formula (14c) applies to correlated errors using congeneric measures then it may be specialized for the parallel measures case, e.g., if y_{12} and y_{13} have nonindependent errors and y_{12} and y_{23} have independent errors:

(a) $\hat{A}^*_{13} = \hat{A}^*_{23}$, by parallel test assumptions, therefore $\hat{A}^*_{12}\hat{A}^*_{13}\hat{\rho}_{T_2T_3} = \hat{A}^*_{12}\hat{A}^*_{23}\hat{\rho}_{T_2T_3}$,

(b) but $\hat{\rho}_{(y_{12},\, y_{23})} = \hat{A}^*_{12}\hat{A}^*_{23}\hat{\rho}_{T_2T_3}$.

Since

$$\hat{A}^*_{12} = \sqrt{\hat{\rho}(y_{12},\, y_{22})} \ , \quad \hat{A}^*_{13} = \sqrt{\hat{\rho}(y_{13},\, y_{23})} \ , \quad \hat{V}_{y_{12}} = \hat{V}_{y_{22}} \ , \quad \hat{V}_{y_{13}} = \hat{V}_{y_{23}} \ ,$$

equation (14c) becomes

$$\hat{\rho}_{\Delta} = \frac{\hat{V}_{y_{12}}\hat{\rho}(y_{12},\, y_{22}) + \hat{V}_{y_{13}}\hat{\rho}(y_{13},\, y_{23}) - 2\hat{\rho}(y_{12},\, y_{23})\sqrt{\hat{V}_{12}\hat{V}_{13}}}{\hat{V}_{y_{12}} + \hat{V}_{y_{13}} - 2\hat{\rho}(y_{12},\, y_{13})\sqrt{\hat{V}_{12}\hat{V}_{13}}} \quad . \quad (14d)$$

Equation (14d) is the formula for the reliability of differences for "linked" (i.e., correlated errors) parallel test measures given by Cronbach and Furby (1970, equation (6)), which can be seen to be the parallel measure specialization of the Werts-Linn equation for nonindependent congeneric measures. Similarly from equations (11a), (11b), (11c), (11d), and (11e) it follows that the estimated correlation of status with gain is:

$$\hat{\rho}_{T_2\Delta} = \frac{\hat{C}_{T_3T_2} - \hat{V}_{T_2}}{\sqrt{\hat{V}_{T_2}(\hat{V}_{T_2} + \hat{V}_{T_3} - 2\hat{C}_{T_2T_3})}} \quad . \quad (15a)$$

In the congeneric case with $A_{12} = A_{13}$, this may be transformed into

$$\hat{\rho}_{T_2\Delta} = \frac{\hat{\rho}_{T_2T_3}\hat{A}^*_{13}\sqrt{\hat{V}_{y_{13}}} - \hat{A}^*_{12}\sqrt{\hat{V}_{y_{12}}}}{\sqrt{\hat{A}^{*2}_{12}\hat{V}_{y_{12}} + \hat{A}^{*2}_{13}\hat{V}_{y_{13}} - 2\hat{\rho}_{T_2T_3}\hat{A}^*_{12}\hat{A}^*_{13}\sqrt{\hat{V}_{y_{12}}\hat{V}_{y_{13}}}}} \, . \qquad (15b)$$

Formula (15b) is the correlation of status with gain given by Werts and Linn (1970a, equation (28)) for the case of congeneric measures and correlated errors, i.e., the formula applies also to the independent error case. In the case of parallel independent measures $\rho_{T_2T_3} = \rho(y_{12}, y_{13}) \div \sqrt{\hat{\rho}(y_{12}, y_{22})\hat{\rho}(y_{13}, y_{23})}$ which when substituted into formula (15b) yields the traditional formula for the correlation of status with gain as applied to the elements of $\hat{\underset{\sim}{\Sigma}}$:

$$\hat{\rho}_{T_2\Delta} = \frac{\hat{\rho}(y_{12},y_{13})\sqrt{\hat{V}_{y_{13}}} - \hat{\rho}(y_{12},y_{22})\sqrt{\hat{V}_{y_{12}}}}{\sqrt{\hat{\rho}(y_{12},y_{22})}\sqrt{\hat{\rho}(y_{12},y_{22})\hat{V}_{y_{12}} + \hat{\rho}(y_{13},y_{23})\hat{V}_{y_{13}} - 2\hat{\rho}(y_{12},y_{13})\sqrt{\hat{V}_{y_{12}}\hat{V}_{y_{13}}}}} \, . \qquad (15c)$$

Our purpose in demonstrating relationships to traditional formulations is purely heuristic, since Jöreskog's program yields estimates of model parameters given the structural assumptions specified by the investigator, i.e., the traditional formulas apply to the elements of $\hat{\underset{\sim}{\Sigma}}$ which are not directly observable but which are estimated as a function of the parameter estimates. Traditional psychometric approaches have dealt with models which are just identified which means that models which exactly reproduce the observed variance-covariance matrix can be employed (i.e., $\underset{\sim}{S} = \hat{\underset{\sim}{\Sigma}}$). The limitation in this approach is that overidentification is necessary if the fit of the model to the data is to be tested.

In this paragraph we propose to use our model to specify the conditions implicit in Cronbach's (1960, pp. 136-139) discussion of coefficients of "stability" and "equivalence." Cronbach uses an example in which two forms of the Mechanical Reasoning Test of the DAT were used, the same forms being used for test and retest purposes. When the same form is repeated, the test-retest correlation is higher than the test-retest correlation between different forms, suggesting the presence of "long-lasting test-specific" factors. The implication is that the errors of measurement for the same test repeated are not independent. Assuming that both forms were repeated and errors of measurement independent for different forms, the model for parallel measures is of the form:

$$\underset{\sim}{y} = \underset{\sim}{\mu} + \underset{\sim}{\Lambda}\underset{\sim}{T} \quad , \qquad (16a)$$

where

$$\underset{\sim}{y} = (y_{12}, y_{22}, y_{13}, y_{23}) \qquad (16b)$$

where y_{12} and y_{13} are the same test as are y_{22} and y_{23} .

$$\underset{\sim}{T} = (T_2, T_3, e_{12}, e_{22}, e_{13}, e_{23}) \quad , \qquad (16c)$$

$$\underset{\sim}{\Lambda} = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix} , \qquad (16d)$$

and

$$\underset{\sim}{\Phi} = \begin{bmatrix} V_{T_2} & & & & & \\ C_{T_2T_3} & V_{T_3} & & \text{Symmetric} & & \\ 0 & 0 & V_{E_{12}} & & & \\ 0 & 0 & 0 & V_{e_{22}} & & \\ 0 & 0 & C_{e_{12}e_{13}} & 0 & V_{e_{13}} & \\ 0 & 0 & 0 & C_{e_{22}e_{23}} & 0 & V_{e_{23}} \end{bmatrix} , \qquad (16e)$$

where $V_{e_{12}} = V_{e_{22}}$, $V_{e_{13}} = V_{e_{23}}$.

The model of (16a) is the special case of factor analysis in which the residual factors are treated as latent factors. Examination of $\underset{\sim}{\Phi}$ shows that the same test errors of measurement are nonindependent, i.e., $C_{e_{12}e_{22}}$ and $C_{e_{22}e_{23}} \neq 0$. All parameters are identifiable and df = 3 (10 distinct elements in $\underset{\sim}{\Sigma}$ less 7 free and constrained parameters). Essentially tau-equivalent assumptions would still have provided identification but with only one overidentifying restriction (since $V_{e_{12}} \neq V_{e_{22}}$, $V_{e_{13}} \neq V_{e_{23}}$) . An interesting case occurs with congeneric assumptions in which case the model is underidentified; however, the unattenuated trait correlation $\rho_{T_2T_3}$ is just identified $[\hat{\rho}^2_{T_2T_3} = \hat{C}(y_{12}, y_{23})\hat{C}(y_{22}, y_{13}) \div \hat{C}(y_{12}, y_{22})\hat{C}(y_{13}, y_{23})]$. Identification may be achieved with the congeneric model by repeating only one test (assuming $A_{12} = A_{13}$) and using different method measures for y_{22} and y_{23} in which case the model is:

$$\underset{\sim}{\Lambda} = \begin{bmatrix} A_{12} & 0 & 1 & 0 & 0 & 0 \\ A_{22} & 0 & 0 & 1 & 0 & 0 \\ 0 & A_{13} & 0 & 0 & 1 & 0 \\ 0 & A_{23} & 0 & 0 & 0 & 1 \end{bmatrix} , \qquad (16f)$$

where $A_{12} = A_{13}$ by assumption, and

$$\Phi = \begin{bmatrix} 1 & & & & & \\ C_{T_2T_3} & V_{T_3} & & \text{Symmetric} & & \\ 0 & 0 & V_{e_{12}} & & & \\ 0 & 0 & 0 & V_{e_{22}} & & \\ 0 & 0 & C_{e_{12}e_{13}} & 0 & V_{e_{13}} & \\ 0 & 0 & 0 & 0 & 0 & V_{e_{23}} \end{bmatrix} . \qquad (16g)$$

This model is just identified (10 distinct elements in $\underset{\sim}{\Sigma}$ less 10 parameters to be estimated). Let us return to Cronbach's example in which there are Forms A (y_{12}) and B (y_{22}) initially and retests on Forms A′ (y_{23}) and B (y_{23}) three years later. Cronbach partitions the variance using the immediate and retest correlations among forms (assumed parallel) which in our model corresponds to the elements of $\underset{\sim}{\Sigma}$. We may translate Cronbach's partitioning procedure into functions of the model parameters in equations (16a), (16b), (16c), (16d), and (16e) as follows:

1. "Lasting General Variance" = $\rho(y_{12}, y_{23}) = A^*_{12}\rho(T_2, T_3)A^*_{23}$ which according to the model equals $\rho(y_{22}, y_{13}) = A^*_{22}\rho(T_2, T_3)A^*_{13}$.
2. "Temporary General Variance" = $\rho(y_{12}, y_{22}) - \rho(y_{12}, y_{23}) = A^*_{12}A^*_{22} - A^*_{12}\rho(T_2, T_2)A^*_{23}$ which according to the model equals $\rho(y_{12}, y_{22}) - \rho(y_{22}, y_{13}) = A^*_{12}A^*_{22} - A^*_{22}\rho(T_2, T_3)A^*_{13}$. In principle there is a different "Temporary General Variance" for the end time $\rho(y_{13}, y_{23}) - \rho(y_{12}, y_{23}) = A^*_{13}A^*_{23} - A^*_{12}\rho(T_2, T_3)A^*_{23}$ which equals $\rho(y_{13}, y_{23}) - \rho(y_{22}, y_{13}) = A^*_{13}A^*_{23} - A^*_{22}\rho(T_2, T_3)A^*_{13}$.
3. "Lasting Specific Variance" for Form A $\rho(y_{12}, y_{13}) - \rho(y_{12}, y_{23}) = \rho(y_{12}, y_{13}) - \rho(y_{22}, y_{13}) = \sqrt{1 - (A^*_{12})^2}\ \rho(e_{12}, e_{13})\ \sqrt{1 - (A^*_{13})^2}$ and

for Form B $\rho(y_{22}, y_{23}) - \rho(y_{12}, y_{23}) = \rho(y_{22}, y_{23}) - \rho(y_{22}, y_{13}) =$
$\sqrt{1 - (A^*_{22})^2}\, \rho(e_{22}, e_{23}) \sqrt{1 - (A^*_{23})^2}$.

4. "Temporary Specific Variance" $[1 - \rho(y_{12}, y_{22})] - [\rho(y_{12}, y_{13}) - \rho(y_{12}, y_{23})] = [1 - \rho(y_{12}, y_{22})] - [\rho(y_{12}, y_{13}) - \rho(y_{22}, y_{13})] =$ $1 - A^*_{12}A^*_{22} - \sqrt{1 - (A^*_{12})^2}\, \rho(e_{12}, e_{13}) \sqrt{1 - (A^*_{13})^2}$ for the correlations used by Cronbach, but in principle there are three other temporary specific variances $1 - A^*_{12}A^*_{22} - \sqrt{1 - (A^*_{22})^2}\, \rho(e_{22}, e_{23}) \sqrt{1 - (A^*_{23})^2}$, $1 - A^*_{13}A^*_{23} - \sqrt{1 - (A^*_{12})^2}\, \rho(e_{12}, e_{13}) \sqrt{1 - (A^*_{13})^2}$, and $1 - A^*_{13}A^*_{23} - \sqrt{1 - (A^*_{22})^2}\, \rho(e_{22}, e_{23}) \sqrt{1 - (A^*_{23})^2}$.

It can be seen that Cronbach's procedure for partitioning of variance involves complicated functions of the model parameters. Not only is it simpler to analyze observed correlations in terms of a set of structural parameters, but it allows for analysis of overidentified models. Further light can be shed on the assumptions implicit in the model of (16a), (16b), (16c), (16d), and (16e) by asking what variables account for the correlated errors. Assuming that a single factor (M_1) underlies the correlation for Form A and another factor (M_2) for Form B the model becomes:

$$\underset{\sim}{y} = \underset{\sim}{\mu} + \underset{\sim}{\Lambda}\underset{\sim}{T} + \underset{\sim}{e}' \quad , \qquad (17a)$$

where

$$\underset{\sim}{y} = (y_{12}, y_{22}, y_{13}, y_{23}) \quad , \qquad (17b)$$

$$\underset{\sim}{T} = (T_2, T_3, M_1, M_2) \quad , \qquad (17c)$$

$$e' = (e'_{12}, e'_{22}, e'_{13}, e'_{23}) \quad , \qquad (17d)$$

$$\underset{\sim}{\Lambda} = \begin{bmatrix} 1 & 0 & B_{12} & 0 \\ 1 & 0 & 0 & B_{22} \\ 0 & 1 & B_{13} & 0 \\ 0 & 1 & 0 & B_{23} \end{bmatrix} , \qquad (17e)$$

and

$$\Phi = \begin{bmatrix} V_{T_2} & & & \\ C_{T_2T_3} & V_{T_3} & & \\ 0 & 0 & 1 & \\ 0 & 0 & 0 & 1 \end{bmatrix} . \qquad (17f)$$

Analysis of the identification problem shows that B_{12}, B_{22}, B_{13}, and B_{23} are not separately identifiable; only the products $(B_{12}\ B_{13})$ and $(B_{22}\ B_{23})$ are identified. This means that in Jöreskog's program we may arbitrarily set $B_{12} = B_{13}$ and $B_{22} = B_{23}$ without disturbing the estimation for other parameters. Assuming $B_{12} = B_{13}$ and $B_{22} = B_{23}$, this model is a simple transformation of (16a), (16b), (16c), (16d), and (16e) under essentially tau-equivalent assumptions, that is, $V_{e_{12}} \neq V_{e_{22}}$, $V_{e_{13}} \neq V_{e_{23}}$ in equation (16e). In particular it can be seen that it must be assumed that M_1 and M_2 are uncorrelated. It is possible to deal with oblique true and method factors but usually more different method measures are required as in our 3 trait x 3 method example in Section I.

When methods of measuring a trait are made as different as possible, it is usually the case that the units of measurement are different, which means that congeneric rather than essentially tau-equivalent or parallel assumptions are appropriate. Werts and Linn (1970a) consider growth models based on congeneric measures, e.g., in one case they use three congeneric measures of T_2 and two congeneric measures of T_3, allowing for same test correlated errors over time. This model is overidentified, but no attempt was made to deal with this complication. Phrasing this problem in terms of Jöreskog's general model:

$$\underset{\sim}{y} = \underset{\sim}{\mu} + \underset{\sim}{\Lambda}\underset{\sim}{T} + \underset{\sim}{e} \qquad (18a)$$

$$\underset{\sim}{y} = (y_{12}, y_{22}, y_{32}, y_{13}, y_{23}) \qquad (18b)$$

where y_{12} and y_{13} are linked as are y_{22} and y_{23}.

$$\underset{\sim}{T} = (T_2, T_3, M_1, M_2) \qquad (18c)$$

$$\Lambda = \begin{bmatrix} A_{12} & 0 & B_{12} & 0 \\ A_{22} & 0 & 0 & B_{22} \\ A_{32} & 0 & 0 & 0 \\ 0 & A_{13} & B_{13} & 0 \\ 0 & A_{23} & 0 & B_{23} \end{bmatrix} \qquad (18d)$$

$$\Phi = \begin{bmatrix} 1 & & & \\ C_{T_2T_3} & V_{T_3} & & \\ 0 & 0 & 1 & \\ 0 & 0 & 0 & 1 \end{bmatrix} . \qquad (18e)$$

Assuming that $A_{12} = A_{13}$, $A_{22} = A_{23}$ and for convenience that $B_{12} = B_{13}$, $B_{22} = B_{23}$, this model has four overidentifying restrictions (15 distinct elements in $\underset{\sim}{\Sigma}$ less 11 parameters to be estimated). Werts and Linn give two formulas (1970a, p. 198, equations (28) and (29)) for estimating the correlation of status with gain involving observed correlations and variances whereas Jöreskog's approach generates a single estimate by equation (15a). In essence Werts and Linn dealt with the elements of the observed variance-covariance matrix $\underset{\sim}{S}$ which may yield inconsistent estimates of $\rho_{T_2\Delta}$ whereas such inconsistency cannot arise with respect to the elements in $\hat{\underset{\sim}{\Sigma}}$. Jöreskog has an unpublished operating program for estimating factor scores within the confirmatory factor analysis model (Jöreskog, 1971). As Cronbach and Furby (1970) note, however, there is seldom need for such estimates.

Relationship to Factor Analysis

A common practice in the factor analysis of growth data is to compare standardized factor loadings at one time to the loadings for the same set of measures at a later time. If the pattern of loadings remains constant over time the inference is drawn that the factors are measuring essentially the same dimension at different times. For example we might have three measures of T_2 at time 1 with factor loadings $A^*_{12} = .30$, $A^*_{22} = .40$, and $A^*_{32} = .50$ and identical loadings on T_3 when these measures are repeated at time 2, i.e., $A^*_{13} = .30$, $A^*_{23} = .40$, and $A^*_{33} = .50$. For heuristic purposes let us suppose that the repetition of tests did not result in methods factors and that the true variance increased from $V_{T_2} = 1.0$ to $V_{T_3} = 1.5$ over time and $C_{T_2T_3} = 1.2$. It may be immediately inferred that the error variances for all tests increased over time since the test reliabilities (in this model the squared factor loadings) remained constant and the true variance increased. However, Wiley and Wiley (1970) have persuasively argued that it is more likely that error variances are a test characteristic which is likely to remain constant over time. If this is so, then an increase in true variance along the same dimension will necessarily mean that the reliabilities of the tests will increase over time, i.e., the standardized factor loadings will increase. In the same fashion it may be deduced that if for any given test over time the unstandardized regression weights $(A_{i2} = A_{i3})$ and the error variances $(V_{e_{i2}} = V_{e_{i3}})$ are equal, then in general the standardized factor loadings (A^*_{ij}) are not proportional from one time to another. We conclude that comparison of standardized factor loading patterns over time provides no logical base for any conclusions about whether pretests and posttests are measuring the same variable. It appears to us that such an assumption, which in this model is equivalent to equality of unstandardized regression weights over time (e.g., $A_{12} = A_{13}$) , is basically not testable within the framework of this model. It would seem better not to make dubious assumptions that either the reliability or the error variance are relatively constant (over time) test characteristics, but to build models and gather requisite information such that these model parameters are identified.

While it is not possible to test the assumption that $A_{12} = A_{13}$, it is quite possible for this assumption to be incompatible with the assumption that $A_{22} = A_{23}$. The ratio of V_{T_3} to V_{T_2} resulting from $A_{12} = A_{13}$ may differ from the ratio resulting from $A_{22} = A_{23}$. This may be tested by the increase in χ^2 (df = 1) resulting from the addition of $A_{22} = A_{23}$ to the model in which $A_{12} = A_{13}$. Within the framework of this model, if it is true that the corresponding pairs of tests over time in fact have the same units, then the scaling of V_{T_3} to V_{T_2} should be the same for each pair.

The finding that the data are consistent with the hypothesis that $A_{12} = A_{13}$ and $A_{22} = A_{23}$ does not necessarily imply that the units of measurement for the corresponding pairs of tests over time are the same since it is quite possible for the scaling to be erroneous for both pairs of tests but in the same way. If the data are inconsistent with the hypothesis that $A_{12} = A_{13}$ and $A_{22} = A_{23}$ we could conclude that the units over time are not the same for both sets of tests, but it is still possible that the units are the same for one of the sets over time. Even if it could be shown that $A_{12} = A_{13}$, this would only be evidence consistent with, not proof of, the hypothesis that the scales are measuring the same process over time.

Determinants of Growth

We-ts and Linn (1970b) have considered the problem of making inferences about the determinants in a linear model. The Werts-Linn formulation was based on classical true score assumptions, i.e., no provision was made for methods factors. For heuristic purposes let us reconsider the problem of growth determinants, formulating the three trait, three method model in terms of growth $(T_3 = T_2 + \Delta)$:

$$\underset{\sim}{T} = (T_1, T_2, \Delta, M_1, M_2, M_3) \tag{19a}$$

$$\underset{\sim}{B} = \begin{bmatrix} A_{11} & 0 & 0 & B_{11} & 0 & 0 \\ A_{21} & 0 & 0 & 0 & B_{21} & 0 \\ A_{31} & 0 & 0 & 0 & 0 & B_{31} \\ 0 & A_{12} & 0 & B_{12} & 0 & 0 \\ 0 & A_{22} & 0 & 0 & B_{22} & 0 \\ 0 & A_{32} & 0 & 0 & 0 & B_{32} \\ 0 & A_{12} & A_{12} & B_{13} & 0 & 0 \\ 0 & A_{22} & A_{22} & 0 & B_{23} & 0 \\ 0 & A_{32} & A_{32} & 0 & 0 & B_{33} \end{bmatrix} \tag{19b}$$

$$\Phi = \begin{bmatrix} 1 & & & & & \\ C_{T_1T_2} & 1 & & & & \\ C_{T_1\Delta} & C_{T_2\Delta} & V_\Delta & & & \\ C_{T_1M_1} & C_{T_2M_1} & C_{\Delta M_1} & 1 & & \\ C_{T_1M_2} & C_{T_2M_2} & C_{\Delta M_2} & C_{M_1M_2} & 1 & \\ C_{T_1M_3} & C_{T_2M_3} & C_{\Delta M_3} & C_{M_1M_3} & C_{M_2M_3} & 1 \end{bmatrix} . \tag{19c}$$

It should be noted that although this formulation does not directly involve the parameters of the underlying growth model $\Delta = D_{\Delta T_1 . T_2} T_1 + D_{\Delta T_2 . T_1} T_2 + \xi$, however, the regression weights are:

$$D_{\Delta T_1 . T_2} = \frac{C_{T_1\Delta} - C_{T_2\Delta}C_{T_1T_2}}{1 - C^2_{T_1T_2}} , \tag{19d}$$

and

$$D_{\Delta T_2 . T_1} = \frac{C_{T_2\Delta} - C_{T_1\Delta}C_{T_1T_2}}{1 - C^2_{T_1T_2}} . \tag{19e}$$

Traditional test theorists (e.g., Bloom, 1964; Thorndike, 1966) have been very concerned with and have drawn substantive inferences about the determinants of growth from the correlation of status with gain, usually corrected for "attenuation." However, as detailed by Werts and Linn (1970b), in a linear structural model prime interest is in the model parameters $D_{\Delta T_1 \cdot T_2}$ and $D_{\Delta T_2 \cdot T_1}$ since if either one is zero the inference will be drawn that the corresponding variable does not directly influence gain. Except in the case in which initial status is uncorrelated with all determinants of growth, knowledge of the correlation of status with gain, $\rho_{T_2\Delta}$, does not allow us to draw inferences about model parameters. It is quite possible for $\rho_{T_2\Delta}$ to be completely spurious due to a common antecedent influence or it is quite possible for $\rho_{T_2\Delta}$ to be zero without implying that $D_{\Delta T_1 \cdot T_2}$ or $D_{\Delta T_2 \cdot T_1}$ be zero. For this reason we question Thorndike's (1966, p. 124) interpretation: "In considerable part, the factors that produce gains during a specified time span appear to be different from those that produced the level of competence exhibited at the beginning of the period." Our objection is that Thorndike's conclusion was made from the correlation of status with gain, without specifically introducing into the analysis any presumed determinants of growth. In a linear structural model the total association of initial status with growth is an insufficient basis for drawing inferences about the various possible determinants of growth.

Discussion

The variety of test response tendencies covered by the rubric "methods factors" appear to be an almost universal complication in sociopsychological growth studies. Even though in principle the multitrait-multimethod model presented in this paper provides for "methods factors," it does not follow that this model does in fact provide a better simulation of reality than previous models which have typically ignored methods factors by assuming independent errors of measurement. It may be expected that our procedure will typically yield different parameter estimates (e.g., correlation of status with gain) than previous procedures, but what has been learned about growth and its determinants thereby? What is learned about reality from the overwhelming concern of the factor analyst with statistical fit? There is no guarantee that the best fitting model yields substantively meaningful results (e.g., Werts, Jöreskog, & Linn, 1971). Why bother with complicated structural models involving unmeasured variables when it is likely that a simple regression equation involving only measured variables will provide the best prediction of the criterion? From our perspective, if the researcher's basic interest is in reality, then the research must be designed to explore reality, i.e., to offer evidence as to which of the initially plausible alternative hypotheses (models) provides the better simulation. In some cases this may involve a study of the theoretical implications to see what information is necessary to discriminate between the alternative models. In other cases the study may be a continuing one as in the building of models to simulate the national economy, in which case the ability to better predict new yearly data is used to discriminate among models. Our purpose in making these remarks is to heighten the awareness of researchers that parameter estimates, such as the reliability of gain scores, are always made within the framework of a whole set of untested assumptions about the nature of reality. It is misleading to talk about "*the* correlation of status with gain" since the meaning of this parameter is totally a function of the particular model used to derive the parameter. In most cases in which this type of estimate has been used, no effort has been made to examine the validity or even plausibility of the models underlying these estimates. The linear structural model presented herein is as suspect as any other model and needs to be justified as one of the plausible alternative hypotheses, prior to data analysis.

References

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.

Cronbach, L. J. *Essentials of psychological testing*. New York: Harper & Brothers, 1960.

Cronbach, L. J., & Furby, L. How we should measure "change"--or should we? *Psychological Bulletin*, 1970, *74*, 68-80.

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

Hauser, R. M., & Goldberger, A. S. The treatment of unobservable variables in path analysis. Social Systems Research Institute Workshop Series EME7030, University of Wisconsin, August 1970.

Jöreskog, K. G. A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, 1969, *34*, 183-202.

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

Jöreskog, K. G. A general method for estimating a linear structural equation system. Research Bulletin 70-54. Princeton, N. J.: Educational Testing Service, September 1970. (b)

Jöreskog, K. G. Statistical analysis of sets of congeneric tests. *Psychometrika*, 1971, *36*, 109-133.

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

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. New York: Addison-Wesley, 1968.

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

Werts, C. E., & Linn, R. L. Path analysis: psychological examples. *Psychological Bulletin*, 1970, *74*, 193-212. (a)

Werts, C. E., & Linn, R. L. A general linear model for studying growth. *Psychological Bulletin*, 1970, *73*, 17-22. (b)

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

Wiley, D. F., & Wiley, J. A. The estimation of measurement error in panel data. *American Sociological Review*, 1970, *35*, 112-117.

V. Conclusions

Sections III and IV constitute fulfillment of the project objectives as stated in the original proposal. The entire written output of this project has been or is in the process of being disseminated to the various relevant audiences. All material has been published or been accepted for formal publication in the final form given in this report.

The substantive conclusions of this project are stated in sections III and IV. While we have succeeded in integrating the methodological literature within the scope of the project, the limitations of our approach need to be stated. The study of the methodological literature alone cannot lead to any conclusions about which kinds of educational growth problems it would be appropriate to apply these methods to. It is much clearer in the physical sciences that quantitative analysis is appropriate only when the mathematical model underlying the analytical procedure approximately simulates the process under study. In the social sciences it is typically unclear whether the model underlying the statistics being used has any resemblance to the phenomenon, usually because we know very little about how the phenomenon actually works. In our judgment, priority should be given to work that attempts to match methodology to particular substantive problems.

VI. Appendix

1. Comment on "The estimation of measurement error in panel data."

2. Comment on Boyle's "Path analysis and ordinal data."

3. Errata to the Werts-Linn comments on Boyle's "Path analysis and ordinal data."

4. Another perspective on "Linear regression, structural relations, and measurement error."

5. A congeneric model for platonic true scores.

6. Estimating true scores using group membership.

7. Errors of inference due to errors of measurement.

8. Identification and estimation in path analysis with unmeasured variables.

9. Intraclass reliability estimates: Testing structural assumptions.

COMMENT* ON "THE ESTIMATION OF MEASUREMENT ERROR IN PANEL DATA"

Wiley and Wiley (1970) have made a contribution to the literature on dealing with errors of measurement by showing how to build a model employing the assumption of homogeneity of error variance in panel data. They argue that this assumption is more plausible than the assumption that the reliability remains constant over time (Heise, 1969). Since we have available data which allow a statistical test of which assumption is the most plausible, this note was written to give the results of this test and to demonstrate how such tests can be performed when at least four sequential measurements are available.

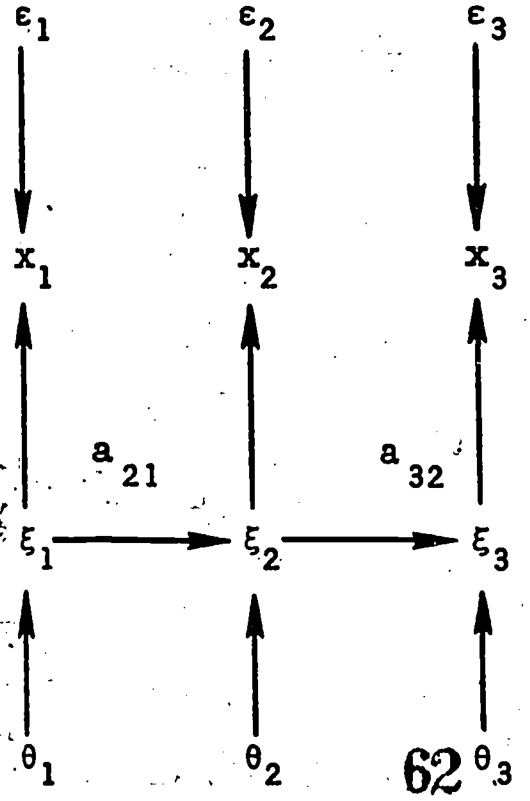


FIGURE 1. A Three Wave Model

The model employed by Wiley and Wiley (1970) is shown in Fig. 1. In this model the reliability of a measure (x_i) is the square of the correlation (ρ_i) between that measure and its underlying true score (ξ_i). Denoting a^*_{21} and a^*_{32} as the standardized path coefficients corresponding to a_{21} and a_{32} respectively, path analysis indicates that the correlations generated by the model are:

$$\begin{aligned} \rho(x_1 x_2) &= \rho_1 a^*_{21} \rho_2 \\ \rho(x_1 x_3) &= \rho_1 a^*_{21} a^*_{32} \rho_3 \\ \rho(x_2 x_3) &= \rho_2 a^*_{32} \rho_3 \end{aligned} \tag{1}$$

It follows from (1) that

$$\rho^2_2 = \frac{\rho(x_1 x_2)\,\rho\,(x_2 x_3)}{\rho(x_1 x_3)} \tag{2}$$

$$[\rho_1 a^*_{21}]^2 = \frac{\rho(x_1 x_2)\,\rho\,(x_1 x_3)}{\rho(x_2 x_3)} \tag{3}$$

$$[\rho_3 a^*_{32}]^2 = \frac{\rho(x_1 x_3)\,\rho\,(x_2 x_3)}{\rho(x_1 x_2)} \tag{4}$$

Thus, without making any assumptions about homogeneity of error variances or reliabilities, it has been demonstrated that the reliability of x_2 (ρ^2_2) is identifiable, and hence also that the corresponding error variance $V(\epsilon_2)=V(x_2)[1-\rho^2_2]$ and true score variance $V(\xi_2)=V(x_2) - V(\epsilon_2)$ is identifiable. For the two outer measures x_1 and x_3, only the products $[\rho_1 a^*_{21}]$ and $[\rho_3 a^*_{32}]$ are identifiable.

Now consider the case in which four sequential measurements are available. Making the same assumptions about the fourth measure that Wiley and Wiley (1970) made about the first three, the model in Fig. 2 is obtained. Generalizing the results of equations (2), (3), and (4), we see that in Fig. 2:

(a) ρ_2, $V(\epsilon_2)$, $V(\xi_2)$, and the product $[\rho_1 a^*_{21}]$ may be identified using either x_1, x_2, and x_3 or x_1, x_2 and x_4

(b) ρ_3, $V(\epsilon_3)$, $V(\xi_3)$, and the product $[\rho_4 a^*_{43}]$ may be identified using either x_1, x_3 and x_4 or x_2, x_3, and x_4.

Path analysis of Fig. 2 also indicates that $\rho(x_2 x_3) = \rho_2 a^*_{32} \rho_3$ and $\rho(x_1 x_4) = \rho_1 a^*_{21} a^*_{32} a^*_{43} \rho_4$, which means that a^*_{32} is overidentified. There-

[*] The research reported herein was performed pursuant to Grant No. OEG–2–700033(509) with the United States Department of Health, Education, and Welfare and the Office of Education.

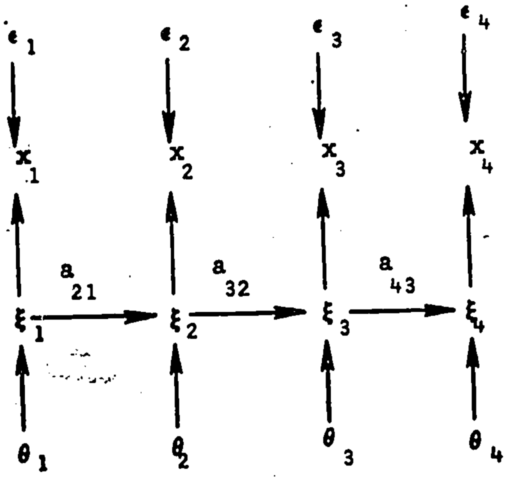


FIGURE 2. A Four-Wave Model

Table 1. Correlations for Quantitative (below Unities) and Verbal (above Unities) Test Scores.[a]

Grade	5	7	9	11
5	1.000	.849	.795	.779
7	.742	1.000	.868	.838
9	.718	.747	1.000	.860
11	.687	.686	.791	1.000

[a] Standard deviations for quantitative scores are 8.986, 13.771, 16.986, and 17.699, respectively; standard deviations for verbal scores are 11.748, 12.704, 13.756, and 14.379, respectively.

fore $a_{32}=a^*_{32}\sqrt{V(\xi_3)\div V(\xi_2)}$ is identifiable. Generalizing to multiple wave panel studies, we may state that, when the assumptions of the Wiley and Wiley structural model are given, error variances, true score variances, and unstandardized regression weights between corresponding true scores are identified for all but the first and last measures. For this reason it appears unnecessary to make either the equal reliability or the equal error variance assumption for inner measures. However, one might wish to know which is the better assumption to make about the first and last measures in order tc identify the corresponding true and error variances and regression weights among true scores. Given at least four-wave data, suggestive but not conclusive evidence about which (if either) assumption is better may be obtained by comparing the estimated error variances and reliabilities for the inner measures.

The four-wave data to be analyzed using the model in Fig. 2 were collected in a longitudinal study (Anderson and Maier, 1963), which in-

Table 2. Model Parameter Estimates and Goodness of Fit Tests.

	Estimates[†]					Fit		
Model	$[\beta_1\hat{a}_{21}]$	β_2	β_3	$[\beta_4\hat{a}_{43}]$	$\hat{a}_{32}$	χ^2	d.f.	p
	SCAT-V Data							
Fig. 2	.884	.960	.942	.912	.959	1.38	1	.240
$\hat{a}_{32} = 1$	.877	.941	.927	.903	1.000	42.61	2	.000
$\rho_2 = \rho_3$	.816	.952	.952	.956	.959	2.17	2	.338
$V(\varepsilon_2) = V(\varepsilon_3)$	.887	.950	.952	.908	.960	12.18	2	.002
	SCAT-Q Data							
Fig. 2	.851	.872	.919	.860	.925	2.78	1	.095
$\hat{a}_{32} = 1$	.823	.840	.899	.852	1.000	42.77	2	.000
$\rho_2 = \rho_3$	.557	.899	.899	.894	.924	5.40	2	.067
$V(\varepsilon_2) = V(\varepsilon_3)$	.851	.873	.918	.861	.925	2.80	2	.247

[†] The symbol "^" denotes an estimate of a population parameter based on sample data.

cluded a group of students tested in the 5th, 7th, 9th and 11th grades with the School and College Ability Tests (SCAT), which yields a Quantitative (Q) and a Verbal (V) score. Table 1 gives (previously unreported) correlations and standard deviations on these tests for a sample of 703 males with complete data.

As Goldberger (1970) notes, the path analysis literature offers no guidance in systematic estimation of overidentified models, such as that depicted in Fig. 2. To obtain estimates, we used Jöreskog's (1970a) general method for the analysis of covariance structures with its associated computer program (Jöreskog *et al.*, 1970). The four-wave model in Fig. 2 is of the *quasi Markov simplex* type, the analysis and programming of which is discussed in detail by Jöreskog (1970b). Under the assumption that the observed distributions are normal (reasonable for these data), Jöreskog's procedure yields maximum likelihood estimates of model parameters and a large sample chi squared test is computed for testing the fit of the model to the data. Furthermore, the program allows certain model parameters to be specified as equal to other parameters or to some constant. This is useful for the present problem because the chi square fit before imposing a restriction (e.g., equal error variances) can be compared to the chi square fit for the more restricted model as a measure of the tenability of that restriction. The analysis proceeded in four steps:

1. The model in Fig. 2 was analyzed without assumptions about equal error variances or reliabilities.

2. To test whether it is reasonable to believe that ξ_2 and ξ_3 are perfectly correlated, the a priori restriction that $a^*_{23} = 1.0$ was imposed. The chi square for this condition less the chi square for the first condition is the chi square with one degree of freedom for testing the restriction.

3. To test the equal reliability assumption, the a priori specification was made that $\rho_2 = \rho_3$. The chi square in this condition less the chi square in the first condition yields a chi square with one degree of freedom for this hypothesis. This assumption is equivalent to the assertion that the error variances are a fixed proportion of the corresponding test variances.

4. To test the equal error variance assumption, the specification was made that $V(\epsilon_2) = V(\epsilon_3)$. The chi square test of this hypothesis is the difference between the chi square for this condition and the one for the first condition and also has one degree of freedom.

The results of the above analysis are shown in Table 2. In step one, for both SCAT-V and SCAT-Q, the χ^2 is small, indicating a good fit. The pattern of the estimates is reasonable in that $\hat{\rho}_2$ and $\hat{\rho}_3$ are approximately equal (published test reliabilities are equal and of the same order of magnitude as these estimates), whereas $[\hat{\rho}_1\hat{a}^*_{21}]$ and $[\hat{\rho}_4\hat{a}^*_{43}]$ are lower, as expected since they are the product of a reliability and a true factor correlation. When the assumption that $a^*_{23} = 1$ is inserted, the χ^2 increased significantly (>40) for both SCAT-V and SCAT-Q. The third step testing the equal reliability assumption yielded a fairly good fit, and the difference χ^2 does not suggest that this hypothesis should be rejected; however $[\hat{\rho}_4\hat{a}^*_{43}]$ appears unreasonable since it is approximately equal to $\hat{\rho}_2$ and $\hat{\rho}_3$. For SCAT-V $[\hat{\rho}_4\hat{a}^*_{43}]$ is slightly larger than $\hat{\rho}_2$ and $\hat{\rho}_3$, which would require $\hat{a}^*_{43}$ to be greater than 1.0 for $\hat{\rho}_4$ to equal $\hat{\rho}_2$ and $\hat{\rho}_3$. In step 4 the difference χ^2 for SCAT-V is statistically significant ($\chi^2_1 = 12.18 - 2.38 = 10.8$) although the absolute magnitude of the difference may not be too important. The step 4 results are more sensible than the step 3 results since $[\hat{\rho}_1\hat{a}^*_{21}]$ and $[\hat{\rho}_4\hat{a}^*_{43}]$ are both less than $\hat{\rho}_2$ and $\hat{\rho}_3$. The step 4 difference χ^2 for SCAT-Q (like step 3) is not statistically significant. Overall, these results suggest that the equal reliability assumption gives a good statistical fit but yields theoretically unreasonable results; whereas the equal error variance assumption may yield poorer fit but estimates which are like the original model of step 1.

CHARLES E. WERTS
KARL G. JÖRESKOG
ROBERT L. LINN

Educational Testing Service
Princeton, N. J.

REFERENCES

Anderson, S. B. and M. H. Maier
1963 "34,000 pupils and how they grew." Journal of Teacher Education, 14:212–216.

Heise, D. R.
1969 "Separating reliability and stability in test-retest correlation." American Sociological Review, 34:93–101.

Goldberger, A. S.
1970 "Econometrics and psychometrics: A survey of communalities." Social Systems Research Institute Workshop Series, EME 7013, University of Wisconsin.

Jöreskog, K. G.
1970a "A general method for analysis of covariance structures." Biometrika, 57:239–251.
1970b "Estimation and testing of simplex models." Research Bulletin 70–42, Educational Testing Service, Princeton, New Jersey.

Jöreskog, K. G., G. T. Gruvaeus, and M. van Thillo

1970 "ACOVS, a general computer program for analysis of covariance structures." Research Bulletin 70-15, Educational Testing Service, Princeton, New Jersey.

Wiley, D. E. and J. A. Wiley

1970 The estimation of measurement error in panel data." American Sociological Review, 35:112-117.

Reprinted from THE AMERICAN JOURNAL OF SOCIOLOGY
Vol. 76, No. 6, May 1971

Printed in U.S.A.

COMMENTS ON BOYLE'S "PATH ANALYSIS AND ORDINAL DATA"[1]

Boyle (1970) has made a significant contribution to the literature in showing how to use dummy variables in path analysis as a device for investigating scale characteristics. However, if path analysis is applied in causal analyses without provision for unmeasured "underlying" variables, there is an implicit assumption that the causative variables are measured without error (i.e., perfect reliability and validity). When each scale unit of an independent variable is treated as a category in Boyle's procedure, no measurement error corresponds to no errors of placement into categories. If there are placement errors, then the observed scale category may not correspond to the "true" scale category, that is, the dummy variable set used by Boyle to code the scale units for an independent variable would correspond to an observed set of fallible variables which are indicators of an underlying set of "true" dummy variables. Figure 1 illustrates the relationships among true and observed dummy variables for a four-unit scale, residual arrows corresponding to errors of placement into that scale category. The number of observed dummy variables (D_a, D_b, D_c) is one less than the number of scale units or categories, and the true dummy variables (T_a, T_b, T_c) are shown as nonindependent because inclusion in one category necessarily involves exclusion from another category. Since dummy variables are dichotomous, the product moment correlations among these variables are ϕ coefficients. Application of path principles to figure 1 shows that the system is underidentified since there are only three correlations among observed variables as compared with nine unknowns (three correlations among errors, three correlations among true dummy variables, and three reliabilities). One solution to the underidentification problem is to use at least three experimentally independent indicators of the independent variable, each of which has the same number of scale categories. For example, in the case of three independent indicators each of which has four levels (i.e., categories), the resulting path diagram would include three "observed" dummy variables (e.g., D_{a1}, D_{a2}, D_{a3}) for each "true" dummy variable (e.g., D_a), the placement errors for a given category on one measure being independent of placement errors in the same or different categories for the other two measures. A path

[1] The research reported herein was performed pursuant to grant no. OEG-2-7000 33(509) with the United States Department of Health, Education, and Welfare and the Office of Education.

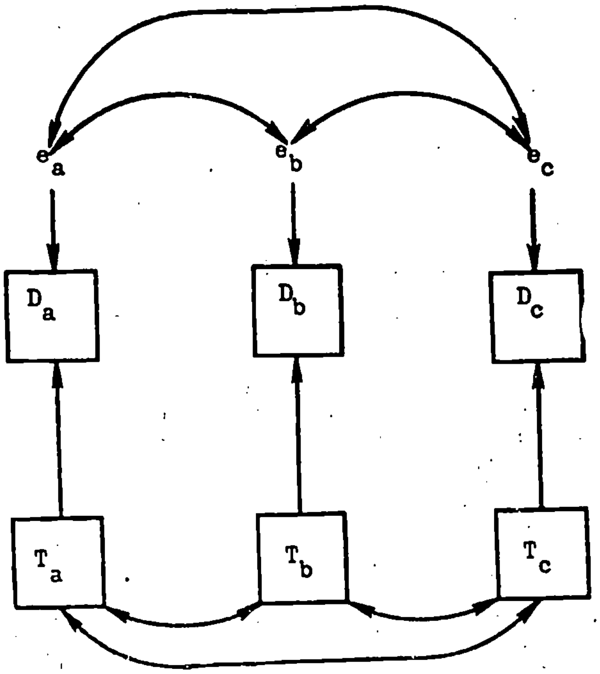


FIG. 1.—Path diagram showing relationships among true and observed dummy variables for a four-unit scale.

analysis of this diagram shows that the system is overidentified (36 observed correlations vs. 21 unknown correlations and path coefficients). When the usual dummy variable coding is used (Decomposition II in Boyle's table 1), the correlation (ϕ) between any two true dummy variables is a function only of the true proportion in these categories:

$$\phi_{T_aT_b} = \sqrt{\frac{P_a}{Q_a}\frac{P_b}{Q_b}}, \tag{1}$$

where $\phi_{T_aT_b}$ is the correlation between T_a and T_b, P_a is the true proportion in category a, P_b is the true proportion in category b, $Q_a = 1 - P_a$, and $Q_b = 1 - P_b$.

It follows from equation (1) that if the correlations among the three true dummy variables are identifiable, then the proportions of the true classification in each category may be identified. The variance of a dichotomous variable is equal to the proportion in that category times the proportion not in that category (e.g., $V_a = P_aQ_a$), and the mean is equal to the proportion in that category (e.g., P_a). The variances and the correlations could be used to calculate covariances or unstandardized regression weights as desired. A dependent variable (Y) may be added to the path diagram, path analysis principles again allowing us to find the equations for the unstandardized regression weights on each of the true dummy variables. When the second type of dummy variable coding in Boyle's table 1 is used, the true regression weights represent the difference between the true Y mean of the group coded "1" in that dummy variable and the true Y mean of the reference group. When Boyle's (1970) first type of dummy variable coding (Decomposition I in table 1 of Boyle's paper) is used, then the true regression weights represent the true difference between successive category means, that is, a test of the equal interval assumption under "effect" scaling. This analysis indicates that one of the reasons that the observed regression weights may differ from one scale category to the next is that the degree of measurement error may differ at different points on the scale.

The analytical model discussed above would still apply if the observations consisted of three independent sorts into a set of nominal categories. In this case the analysis is equivalent to an analysis of variance with fallible group information, and the problem is whether the true group means really differ, that is, whether the regression weights for the true dummy variables are all zero.

In passing it might be noted that for overidentified models of the type discussed above, a procedure for estimating the parameters of the model is needed. As Goldberger (1970, p. 25) notes: "the path analysis literature offers no guidance on systematic estimation of overidentified models." Because the distribution of variables (true and observed) is multinomial, the function to be minimized (Mote and Anderson 1965; Cochran 1968, pp. 647–48) for estimation purposes is a χ^2 involving observed and hypothetical ("expected") probabilities. The dummy variable path analysis equations therefore must be translated into probability functions to obtain estimates in overidentified models. In our opinion, path analysis is useful in this type of problem because it helps deal with questions of identifiability, and it is easier for the researcher to conceptualize the relationships among variables.

Charles E. Werts
Robert L. Linn

Educational Testing Service

REFERENCES

Boyle, R. P. 1970. "Path Analysis and Ordinal Data." *American Journal of Sociology* 75 (January): 461–80.

Cochran, W. G. 1968. "Errors of Measurement in Statistics." *Technometrics* 10: 637–66.

Goldberger, A. A. 1970. *Econometrics and Psychometrics: A Survey of Communalities.*

Social Systems Research Institute Workshop Series EME 7013. Madison: University of Wisconsin.

Mote, V. L., and R. L. Anderson. 1965. An Investigation of the Effect of Misclassification on the Properties of χ^2-Tests in the Analysis of Categorical Data." *Biometrika* 52: 95–109.

Errata to the Werts-Linn Comments on Boyle's "Path Analysis and Ordinal Data"

Charles E. Werts & Robert L. Linn

Educational Testing Service

Princeton, N. J.

Werts & Linn (1971) pointed out that Boyle (1970) had implicitly assumed that the causative variables were measured without error. Further study of literature relating to this problem (e.g., Cochran, 1968; Evans, 1970; Anderson; 1959) indicates that the Werts-Linn procedure for dealing with categorical errors of measurement is incorrect. The purpose of this note is to set the record straight.

As a basis for generalization to polychotomous variables, first consider the case of three independent fallible dichotomous measures X_j ($j = 1, 2, 3$) of an underlying true dichotomy (T). The observed categories will be labelled $k = 1, 2$ and the true categories $\ell = 1, 2$. The relationship between X_j and T can be expressed as a function of the conditional probabilities $P\{X_j = k|T = \ell\} = \theta_{jk\ell}$ for each combination of k and ℓ:

$$P\{X_j = 1|T = 1\} = \theta_{j11}, \; P\{X_j = 1|T = 2\} = \theta_{j12} ,$$

$$P\{X_j = 2|T = 1\} = \theta_{j21}, \text{ and } P\{X_j = 2|T = 2\} = \theta_{j22} .$$

θ_{j21} is commonly labelled the proportion of false negatives and θ_{j12} the proportion of false positives. The sum of the conditional probabilities for a fixed value of ℓ is unity i.e., $\theta_{j11} + \theta_{j21} = 1$ and $\theta_{j12} + \theta_{j22} = 1$. Define $P_{T_\ell} = P\{T = \ell\}$ and $P_{j_k} = P\{X_j = k\}$ where $\overset{\ell}{\Sigma} P_{T_\ell} = 1$ and $\overset{k}{\Sigma} P_{j_k} = 1$. The model parameters to be estimated are the conditional probabilities for each observed measure and the true proportions in each category. Since each

[*] The research reported herein was performed pursuant to Grant No. OEG-2-700033(509) with the United States Department of Health, Education, and Welfare and the Office of Education.

object is categorized by each different measure the proportion of objects for each combination of observed categories can be computed. Define $P_{jk,j'k',j''k''} = P\{X_j = k; X_{j'} = k', X_{j''} = k''\}$ where $j \neq j' \neq j''$.

In the three measure case the observed data consist of eight joint probabilities $P_{11,21,31}$, $P_{11,21,32}$, $P_{11,22,31}$, $P_{11,22,32}$, $P_{12,21,31}$, $P_{12,21,32}$, $P_{12,22,31}$, and $P_{12,22,32}$. The next step is to relate these observed probabilities to the model parameters. Starting with $P_{11,21,31}$ we obtain:

$$\mathcal{E}(P_{11,21,31}) = P\{X_1 = 1, X_2 = 1, X_3 = 1, T = 1\} + P\{X_1 = 1, X_2 = 1, X_3 = 1, T = 2\}$$

Expressed in terms of conditional probabilities the proportions are

$P\{X_1=1,X_2=1,X_3=1,T=1\} = P\{X_1=1|X_2=1,X_3=1,T=1\}P\{X_2=1|X_3= 1,T=1\}P\{X_3=1|T=1\}P\{T=1\}$, and

$P\{X_1=1,X_2=1,X_3=1,T=2\} = P\{X_1=1|X_2=1,X_3=1,T=2\}P\{X_2=1|X_3 =1,T=2\}P\{X_3=1|T=2\}P\{T=2\}$.

The assumption that the measures are independent implies that

$P\{X_1 = 1|X_2 = 1, X_3 = 1, T = 1\} = P\{X_1 = 1|T = 1\} = \theta_{111}$,

$P\{X_2 = 1|X_3 = 1, T = 1\} = P\{X_2 = 1|T = 1\} = \theta_{211}$,

$P\{X_1 = 1|X_2 = 1, X_3 = 1, T = 2\} = P\{X_1 = 1|T = 2\} = \theta_{112}$, and

$P\{X_2 = 1|X_3 = 1, T = 2\} = P\{X_2 = 1|T = 2\} = \theta_{212}$.

Thus, by substitution:

$$\mathcal{E}(P_{11,21,31}) = \theta_{111}\,\theta_{211}\,\theta_{311}\,P_{T_1} + \theta_{112}\,\theta_{212}\,\theta_{312}\,P_{T_2}. \qquad (1)$$

While this process could be repeated for each of the observed joint probabilities, for identification purposes it is better to replace these by the following set:

$P_{11,21} = P_{11,21,31} + P_{11,21,32}$,

$P_{11,31} = P_{11,21,31} + P_{11,22,31}$,

$P_{21,31} = P_{11,21,31} + P_{12,21,31}$,

$P_{11} \quad = P_{11,21,31} + P_{11,21,32} + P_{11,22,31} + P_{11,22,32}$,

$P_{21} \quad = P_{11,21,31} + P_{11,21,32} + P_{12,21,31} + P_{12,21,32}$, and

$P_{31} \quad = P_{11,21,31} + P_{11,22,31} + P_{12,21,31} + P_{12,22,31}$.

Following the procedure used for $P_{11,21,31}$ it may be shown that:

$$\mathcal{E}(P_{11,21}) = \theta_{111}\ \theta_{211}\ P_{T_1} + \theta_{112}\ \theta_{212}\ P_{T_2}, \tag{2}$$

$$\mathcal{E}(P_{11,31}) = \theta_{111}\ \theta_{311}\ P_{T_1} + \theta_{112}\ \theta_{312}\ P_{T_2}, \tag{3}$$

$$\mathcal{E}(P_{21,31}) = \theta_{211}\ \theta_{311}\ P_{T_1} + \theta_{212}\ \theta_{312}\ P_{T_2}, \tag{4}$$

$$\mathcal{E}(P_{11}) = \theta_{111}\ P_{T_1} + \theta_{112}\ P_{T_2}, \tag{5}$$

$$\mathcal{E}(P_{21}) = \theta_{211}\ P_{T_1} + \theta_{212}\ P_{T_2}, \text{ and} \tag{6}$$

$$\mathcal{E}(P_{31}) = \theta_{311}\ P_{T_1} + \theta_{312}\ P_{T_2}. \tag{7}$$

Note that even though we started with eight joint probabilities, we have only seven equations because of the condition that all the observed probabilities sum to unity. If the model parameters are identifiable then it should be possible to solve these equations for each parameter in terms of the expected probabilities. For this purpose it is convenient to define:

$$C_{jk,j'k'} = \mathcal{E}(P_{jk,j'k'}) - \mathcal{E}(P_{jk})\,\mathcal{E}(P_{j'k'}),$$

$$C_{jk,j'k',j''k''} = \mathcal{E}(P_{jk,j'k',j''k''}) - [\mathcal{E}(P_{jk})]\ C_{j'k',j''k''} - [\mathcal{E}(P_{j'k'})]C_{jk,j''k''}$$

$$- [\mathcal{E}(P_{j''k''})]\ C_{jk,j'k'} - \mathcal{E}(P_{jk})\,\mathcal{E}(P_{j'k'})\,\mathcal{E}(P_{j''k''}), \text{ and } Q_{T_\ell} = 1 - P_{T_\ell}.$$

For the dichotomous case $Q_{T_1} = P_{T_2}$. Solving equations 1 through 7 for P_{T_1} we obtain:

$$\frac{Q_{T_1}^2 - P_{T_1}^2}{\sqrt{P_{T_1} Q_{T_1}}} = \frac{C_{11,21,31}}{\sqrt{C_{11,21}\ C_{11,31}\ C_{21,31}}} \tag{8}$$

Equation (8) shows that P_{T_1} and $P_{T_2} = 1 - P_T$ are identified. Further analysis yields:

$$\theta_{112} = \mathcal{E}(P_{11}) - \sqrt{\frac{C_{11,21}\ C_{11,31}}{C_{21,31}} \left(\frac{P_{T_1}}{Q_{T_1}}\right)}\ , \qquad (9)$$

$$\theta_{212} = \mathcal{E}(P_{11}) - \sqrt{\frac{C_{11,21}\ C_{21,31}}{C_{11,31}} \left(\frac{P_{T_1}}{Q_{T_1}}\right)}\ , \qquad (10)$$

and $$\theta_{312} = \mathcal{E}(P_{11}) - \sqrt{\frac{C_{11,31}\ C_{21,31}}{C_{11,21}} \left(\frac{P_{T_1}}{Q_{T_1}}\right)}\ . \qquad (11)$$

Since P_{T_1} and P_{T_2} are identified, equations (9), (10), and (11) show that θ_{112}, θ_{212}, θ_{312} and therefore $\theta_{122} = 1 - \theta_{112}$, θ_{222}, $= 1 - \theta_{212}$, and $\theta_{322} = 1 - \theta_{312}$ are identifiable. Given P_{T_1}, P_{T_2}, θ_{112}, θ_{212}, and θ_{312} identified, equations (5), (6), and (7) show that θ_{111}, θ_{211} and θ_{311} and therefore $\theta_{121} = 1 - \theta_{111}$, θ_{221}, $= 1 - \theta_{211}$, and $\theta_{321} = 1 - \theta_{311}$ are identifiable. Since the model consists of seven equations in seven unknowns (i.e., just identified), parameter estimates can be obtained which will exactly reproduce the observed probabilities, i.e., the observed joint, probabilities would equal the expected joint probabilities estimated from the parameter estimates. The above analysis shows that the true proportions may be identified given three independent dichotomous measures, a point which Werts & Linn (1971) failed to discover. The right side numerator of equation (8) is the expected value of the triple covariance between X_1, X_2, and X_3; which is the crucial piece of information neglected in the Werts-Linn path approach. Furthermore, path analysis usually ignores variable means, which would result in neglect of equations (9), (10), and (11) which involve means (P_{jk}).

Next consider the trichotomous case in which $k = 1,2,3$, $\ell=1,2,3$ and $j=1,2,3$ given the assumption of independent measures. The relationship between the j^{th} observed trichotomy and the true trichotomy involves nine conditional probabilities: θ^*_{j11}, θ^*_{j12}, θ^*_{j21} and θ^*_{j22} as defined previously plus

$P\{X_j = 1|T = 3\} = \theta^*_{j13}$, $P\{X_j = 2|T = 3\} = \theta^*_{j23}$,

$P\{X_j = 3|T = 1\} = \theta^*_{j31}$, $P\{X_j = 3|T = 2\} = \theta^*_{j32}$, and $P\{X_j = 3|T = 3\} = \theta^*_{j33}$.

By definition: $\theta^*_{j11}, \theta^*_{j21}, + \theta^*_{j31} = \theta^*_{j12} + \theta^*_{j22} + \theta^*_{j32} = \theta^*_{j13} + \theta^*_{j23} + \Theta^*_{j33} = 1$.

Let K = total number of categories and J = total number of indpendent measures. The observed data consist of the $K^J=27$ joint triple probabilities $P_{1k,2k',3k''}$, one of which may be expressed as a function of the other 26.

There are $JK^2 = 27$ $\theta^*_{jk\ell}$, JK of which can be stated as a function of the others because for a fixed ℓ the $\theta^*_{jk\ell}$ sum to unity and K =3 P_{T_ℓ} one of which it can be stated as 1 minus the sum of the others.

Therefore there are a total of $JK(K-1) + (K-1) = 20$ independent parameters to be estimated from the $K^J - 1 = 26$ independent observed joint probabilities, i.e., the model has six overidentifying restrictions. This does not necessarily mean that all parameters are identified and in principle the expected value of each $P^*_{jk,j'k',j''k''}$ should be derived as done previously and the equations solved for each parameter. Rather than attempt this directly, it can be seen that if category three were collapsed into category 2 then the analysis would be identical to that shown for dichotomous variables. The relationships would be (*refers to probabilities prior to collapsing categories):

$$P_{T_1} = P^*_{T_1} \quad , \qquad \text{12a}$$

$$\theta_{j11} = \theta^*_{j11} \quad , \qquad \text{12b}$$

$$\text{and} \quad \theta_{j12} \quad (1-P_{T_1}) = \theta^*_{j12} \, P^*_{T_2} + \theta^*_{j13} \, P^*_{T_3} \quad . \qquad \text{12c}$$

From our previous analysis we know that the parameters in the right side of equations 12a,b, & c can be identified from

$P_{11,21,31} = P^*_{11,21,31}$,

$P_{11,21,32} = P^*_{11,21,32} + P^*_{11,21,33}$,

$P_{11,22,31} = P^*_{11,22,31} + P^*_{11,23,31}$,

$P_{11,22,32} = P^*_{11,22,32} + P^*_{11,23,32} + P^*_{11,22,33} + P^*_{11,23,33}$,

$P_{12,21,31} = P^*_{12,21,31} + P^*_{13,21,31}$,

$P^*_{12,21,32} = P^*_{12,21,32}, + P^*_{12,21,33} + P^*_{13,21,32} + P^*_{13,21,33}$,

$$P_{12,22,31} = P^*_{12,22,31} + P^*_{12,23,31} + P^*_{13,22,31} + P^*_{13,23,31}, \text{ and}$$

$$P_{12,22,32} = P^*_{12,22,32} + P^*_{12,22,33} + P^*_{12,23,32} + P^*_{12,23,33} \quad P^*_{13,22,32} +$$

$$P^*_{13,22,33} + P^*_{13,23,32} + P^*_{13,23,33}.$$

These eight $P_{jk,j'k',j''k''}$ could be entered into the analysis shown for dichotomies and the corresponding parameters in 12a,b, & c identified. In a similar fashion if we collapse category 1 into 3 then:

$$P_{T_2} \; P^*_{T_2}, \qquad \text{12d}$$

$$\theta_{j22} = \theta^*_{j22}, \qquad \text{12e}$$

$$\text{and } \theta_{j23} \, (1-P_{T_2}) = \theta^*_{j21} \, P^*_{T_1} + \theta^*_{j23} \, P^*_{T_3}. \qquad \text{12f}$$

The right hand parameters in 12d,e, & f would be identified from:

$$P_{12,22,32} = P^*_{12,22,32},$$

$$P_{12,22,33} = P^*_{12,22,31} + P^*_{12,22,33},$$

$$P_{12,23,32} = P^*_{12,23,32} + P_{12,21,32},$$

$$P_{12,23,33} = P^*_{12,23,33} + P^*_{12,21,33} + P^*_{12,23,31} + P^*_{12,21,31},$$

$$P_{13,22,32} = P^*_{13,22,32} + P^*_{11,22,32},$$

$$P_{13,22,33} = P^*_{13,22,33} + P^*_{13,22,31} + P^*_{11,22,33} + P^*_{11,22,31},$$

$$P_{13,23,32} = P^*_{13,23,32} + P^*_{13,21,32} + P^*_{11,23,32} + P^*_{11,21,32}, \text{ and}$$

$$P_{13,23,33} = P^*_{13,23,33} + P^*_{13,23,31} + P^*_{13,21,33} + P^*_{13,21,31} \quad P^*_{11,23,33} +$$

$$P^*_{11,23,31} + P^*_{11,21,33} + P^*_{11,21,31}.$$

These eight $P_{jk,j'k',j''k''}$ could likewise be entered into the analysis for dichotomies where the two categories are $k = 2,3$ instead of $k = 1,2$ as shown in our original analysis. We can conclude that $P^*_{T_1}$, $P^*_{T_2}$, and $P^*_{T_3}$

are identified from 12a,d. and θ^*_{111}, θ^*_{211}, θ^*_{311}, θ^*_{122}, θ^*_{222}, and θ^*_{322} from equations 12b,e. The remaining 12 parameters in equations 12c and f have six conditions imposed by equations 12c,f so we need six more equations for identification. The simplest set, which is independent of information used in the dichotomous analyses is:

$$\begin{aligned}
P_{11,22} &= P^*_{11,22,31} + P^*_{11,22,32} + P^*_{11,22,33},\\
P^*_{11,32} &= P^*_{11,21,32} + P^*_{11,22,32} + P^*_{11,23,32},\\
P^*_{12,21} &= P^*_{12,21,31} + P^*_{12,21,32} + P^*_{12,21,33},\\
P^*_{12,31} &= P^*_{12,21,31} + P^*_{12,22,31} + P^*_{12,23,31},\\
P^*_{21,32} &= P^*_{11,21,32} + P^*_{12,21,32} + P^*_{13,21,32},\\
\text{and}\quad P^*_{22,31} &= P^*_{11,22,31} + P^*_{12,22,31} + P^*_{13,22,31}.
\end{aligned}$$

Application of the procedure used to derive equation (1) yields:

$$\left.\begin{aligned}
\mathcal{E}(P^*_{11,22}) &= \theta^*_{111}\theta^*_{221}P^*_{T_1} + \theta^*_{112}\theta^*_{222}P^*_{T_2} + \theta^*_{113}\theta^*_{223}P^*_{T_3},\\
\mathcal{E}(P^*_{11,32}) &= \theta^*_{111}\theta^*_{321}P^*_{T_1} + \theta^*_{112}\theta^*_{322}P^*_{T_2} + \theta^*_{113}\theta^*_{323}P^*_{T_3},\\
\mathcal{E}(P^*_{12,21}) &= \theta^*_{121}\theta^*_{211}P^*_{T_1} + \theta^*_{122}\theta^*_{212}P^*_{T_2} + \theta^*_{123}\theta^*_{213}P^*_{T_3},\\
\mathcal{E}(P^*_{12,31}) &= \theta^*_{121}\theta^*_{311}P^*_{T_1} + \theta^*_{122}\theta^*_{312}P^*_{T_2} + \theta^*_{123}\theta^*_{313}P^*_{T_3},\\
\mathcal{E}(P^*_{21,32}) &= \theta^*_{211}\theta^*_{321}P^*_{T_1} + \theta^*_{212}\theta^*_{322}P^*_{T_2} + \theta^*_{213}\theta^*_{323}P^*_{T_3},\\
\text{and}\quad \mathcal{E}(P^*_{22,31}) &= \theta^*_{221}\theta^*_{311}P^*_{T_1} + \theta^*_{222}\theta^*_{312}P^*_{T_2} + \theta^*_{223}\theta^*_{313}P^*_{T_3}.
\end{aligned}\right\}\quad (13)$$

Equations (13) in combination with previously identified parameters and equations (11) and (12c,f)identify the remaining parameters. Note that six equations have not been used, these representing the six degrees of overidentification. The method which appears appropriate for estimating parameters when the observed variables are independent polychotomous measures is discussed in Anderson (1959, sec. 3.6) and Cochran (1968, sec. 6). In this procedure a chi square function involving the observed and estimated expected joint probabilities is minimized as a function of the model parameters. The resulting χ^2 with degrees of freedom equal to the number of overidentifying restrictions,is a measure of the fit of the model to the data. Our analysis indicates that given three independent polychotomies (K = 3) all model parameters are identifiable. The number of overidentifying restrictions is equal to $(K^J - 1) - (JK + 1)(K - 1)$ where J = the number of independent measures and K = the number of categories.

We may now consider exactly why the Werts-Linn analysis was inappropriate to the problem. For this purpose it is helpful to put the conditional probabilities into matrix form where columns refer to observed categories and rows to true categories:

$$\theta_{\sim j}^{*} = \begin{bmatrix} \theta_{j11}^{*} & \theta_{j21}^{*} & \theta_{j31}^{*} \\ \theta_{j12}^{*} & \theta_{j22}^{*} & \theta_{j32}^{*} \\ \theta_{j13}^{*} & \theta_{j23}^{*} & \theta_{j33}^{*} \end{bmatrix}. \qquad (14)$$

As noted earlier each row in $\theta_{\sim j}^{*}$ sums to unity. If the true categories 1, 2, and 3 actually form an ordered set of classifications such that category 1 is "closer" to 2 than to 3 then we would expect that classificatory errors would be more likely for neighboring categories, i.e., $\theta_{j12}^{*} > \theta_{j13}^{*}$ and $\theta_{j32}^{*} > \theta_{j31}^{*}$. In contrast, if the true categories are basically unordered, it would be more reasonable to expect the likelihood of misclassification to be similar for any of the other classes, i.e., $\theta_{j12}^{*} \cong \theta_{j13}^{*}$, $\theta_{j21}^{*} \cong \theta_{j23}^{*}$, and $\theta_{j31}^{*} \cong \theta_{j33}^{*}$. In other words the probability of misclassification is a function of the underlying scale or "true" category in the case of ordered categories and is not in the case of unordered categories. Werts & Linn implicitly assumed that the errors for one category were uncorrelated with the underlying "true" dummy variable for the same and for other categories which translated into the present framework corresponds to the analysis for an unordered scale i.e., for an ordered scale the errors would be correlated with the "true" dummy variables for other categories. It can be algebraically shown that the Werts-Linn procedure leads to incorrect formulae for the expected value of the observed joint probabilities when the categories are ordered. Since Boyle (1970) was examining the problem of ordered categories (i.e., scales) the Werts-Linn approach is not relevant to his problem.

Bibliography

Anderson, T. W.

1959 "Some Scaling Models and Estimation Procedures in the Latent Class Model." In O. Grenander (Ed.), _Probability and Statistics, The Harold Cramer Volume._ New York: Wiley, 1959, pp. 9-38.

Boyle, Richard P.

1970 "Path Analysis and Ordinal Data." _American Journal of Sociology,_ 75 (January) :461-480.

Cochran, William G.

1968 "Errors of Measurement in Statistics." _Technometrics_ 10 (November): 637-666.

Evans, Glen T.

1970 "The Analysis of Categorizing Behavior." _Psychometrika_ 35 (September): 367-392.

Hauser, Robert M., & Goldberger, A. S.

1970 "The Treatment of Unobservable Variables in Path Analysis." Social Systems Research Institute Series EME 7030, University of Wisconsin, Madison, Wisconsin.

Werts, Charles E., & Linn, R. L.

1971 "Comment on 'Path Analysis and Ordinal Data'." _American Journal of Sociology_ 76 (May): in press.

Another perspective on "Linear regression, structural relations, and measurement error."

Charles E. Werts, Robert L. Linn, and

Karl G. Jöreskog

Abstract

A stochastic disturbance term appears to be essential for structural models in the social sciences. The analysis of such models is considered from the perspective of Jöreskog's (1970) general model for the analysis of covariance structure.

Another perspective on "Linear regression, structural relations, and measurement error."

Charles E. Werts, Robert L. Linn, and

Karl G. Jöreskog †

Isaac (1970) has performed a useful service in dispelling the common misconception that parameters estimated in a regression analysis are necessarily those involved in a structural relation. Researchers who would use the formulae supplied by Isaac should be warned, however, that these apply to a model which is seldom, if ever, relevant. Johnston (1963, pg. 148) notes that this model "hardly seems appropriate for econometric work, since, if it were true, the only reason for points not lying exactly on a straight line would be errors of observation. A stochastic component of behavior would seem an essential in economics." This comment applies equally to psychology, in which the usual type of relationship is like that between fathers and sons height, where even if there were no errors of measurement the correlation would be less than perfect. Adding a stochastic disturbance term, μ , the model becomes $Y = \alpha + \beta X + \mu$. Rather than review the analysis of this model, which is covered by Johnston (1963, Chap. 6), we propose to consider the problem from the perspective of Jöreskog's (1970) general model for the analysis of covariance structures.

Jöreskog (1970, pg. 239) considers

> a data matrix $\mathbf{X} = \{x_{\alpha i}\}$ of N observations on p response variables and the following model. Rows of $\mathbf{X}$ are independently distributed, each having a multivariate normal distribution with the same variance-covariance matrix Σ of the form
>
> $$\Sigma = \mathrm{B}(\Lambda\Phi\Lambda' + \Psi^2)\mathrm{B}' + \Theta^2, \qquad (1)$$
>
> and mean vectors given by
>
> $$\mathcal{E}(\mathbf{X}) = \mathrm{A}\Xi\mathrm{P}, \qquad (2)$$
>
> where $\mathrm{A} = \{a_{\alpha s}\}$ is an $N \times g$ matrix of rank g and $\mathrm{P} = \{p_{tj}\}$ is a $h \times p$ matrix of rank h, both being fixed matrices with $g \leqslant N$ and $h \leqslant p$; $\Xi = \{\xi_{st}\}$, $\mathrm{B} = \{\beta_{ik}\}$, $\Lambda = \{\lambda_{km}\}$, the symmetric matrix $\Phi = \{\phi_{mn}\}$, and the diagonal matrices $\Psi = \{\delta_{kl}\psi_k\}$ and $\Theta = \{\delta_{ij}\theta_i\}$ are parameter matrices.

Means, variances and covariances are structured in terms of the parameters in $\Xi, B, \Lambda, \Phi, \Psi$ and Θ which may be (a) "fixed" parameters that have been assigned given values, (b) "constrained" parameters that are unknown but equal to one or more other parameters, and (c) "free" parameters that are unknown and unconstrained.

For analytical purposes let us start with a stochastic disturbance term and errors of measurement as given by Isaac (1970, pg. 214). In this model the observed variables (lower case letters) are $x = X + \varepsilon_x$ and $y = \alpha + \beta X + \mu + \varepsilon_y$. In this problem the question of means is not important (since α can be estimated from the β estimate, $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$) and we may proceed by considering the structure of the variance-covariance matrix of the observed variables. The observed vector is (y, x), the factors are $(X, \mu, \varepsilon_y, \varepsilon_x)$, B is an identity matrix, Ψ and $\Theta = 0$,

$$\Lambda = \begin{bmatrix} \beta & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix},$$

and $$\Phi = \begin{bmatrix} \sigma^2_X & 0 & 0 & 0 \\ 0 & \sigma^2_\mu & 0 & 0 \\ 0 & 0 & \sigma^2_{\varepsilon y} & 0 \\ 0 & 0 & 0 & \sigma^2_{\varepsilon_x} \end{bmatrix}.$$

Since there are 2 x 3 ÷ 2 distinct elements in Σ and five free parameters, this model is underidentified by 2 restrictions (d.f. = 3-5). If the error variances $\sigma^2_{\epsilon_y}$ and $\sigma^2_{\epsilon_x}$ were known apriori (possibly computed from known reliabilities for measures), then the model would be just identified and the associated computer program (Jöreskog, Gruvaes, and von Thillo, 1970) could be used to obtain maximum likelihood estimates of parameters. Because the model is just identified, the estimated elements of $\hat{\Sigma}$ would exactly equal corresponding elements in the observed variance-covariance matrix. Isaac's model involves the deletion of μ, i.e., the second column in Λ and Φ, in which case there are still 3 distinct elements in Σ but the number of free parameters has been reduced to four (β, σ^2_X, $\sigma^2_{\epsilon_y}$, $\sigma^2_{\epsilon_x}$) so that only one additional assumption is needed for identification. If as in Isaac's case #1, $\sigma^2_{\epsilon_x}$ is known, then all parameters are identified. When the ratio of the error variances λ is known (Isaac's case #3) then Ψ and Θ = 0 as before, but now:

$$B^* = \begin{bmatrix} \beta & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix},$$

$$\Lambda^* = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \sqrt{\lambda} & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

$$\text{and } \Phi^* = \begin{bmatrix} \sigma^2_X & 0 & 0 \\ 0 & \sigma^2_{\epsilon_x} & 0 \\ 0 & 0 & \sigma^2_{\epsilon_x} \end{bmatrix}$$

This model has two elements in Φ constrained to be equal and two free parameters (since $\sqrt{\lambda}$ is "fixed") and the model is just identified.[1] Isaac's fourth case, in which $\sigma^2_{\varepsilon_x}$ and $\sigma^2_{\varepsilon_y}$ are both known, is of interest because when these are inserted in Φ the model has one overidentifying restriction. Assuming that the observed distributions are normal, a chi-square with one degree of freedom is generated which tests the fit of the model to the data. In general, Isaac's equation (3) is not the maximum likelihood solution for this overidentified model; this difference arising because equation (3) uses only the ratio of the error variances, neglecting the absolute values.

Because most effects have multiple causes, it is of interest to consider the case of an exact functional model in which there are only three variables X, Y, and Z and causation may occur in any direction. With any variable held constant the true correlation between the other two is perfect, i.e., the true partial correlation between any two variables with the third controlled is unity. However, the partial correlation is equal to the product of the two corresponding partial regression weights, e.g., $\rho_{XY.Z} = \beta_{XY.Z}\,\beta_{YX.Z} = 1.0$ in this model. Therefore, the partial regression weight in one direction is the inverse of that in the opposite direction with the same variable controlled, e.g., $\beta_{XY.Z} = 1/\beta_{YX.Z}$. In the model $Y = \alpha + \beta X + \mu$, the stochastic term represents the effects of all other influences which are assumed to be independent of X. The partial correlation $\rho_{XY.\mu}$ is equal to

$$\rho_{XY.\mu} = \frac{\rho_{XY} - \rho_{X\mu}\,\rho_{Y\mu}}{\sqrt{(1-\rho^2_{X\mu})\,(1-\rho^2_{Y\mu})}} = 1,$$

since $\rho_{X\mu} = 0$, and $\rho^2_{Y\mu} = 1-\rho^2_{XY}$.

Therefore the reciprocal relationship $\beta_{XY.\mu} = 1/\beta_{YX.\mu}$ holds in the stochastic disturbance term model. Since μ is independent of X, $\beta_{YX.\mu} = \beta_{YX}$, but of course since μ is not independent of Y this relationship holds only for Y on X.

A variety of other solutions to the identification problem may be used instead of or in combination with those discussed by Isaac. For example, if a "congeneric" measure (Jöreskog, 1970, sec. 2.2) x_1 of X ($x_1 = \beta_{x_1X} X + e_{x_1}$) were added to the model with the stochastic disturbance term, then β, β_{x_1X}, σ^2_X, $\sigma^2_{\epsilon_x}$, $\sigma^2_{\epsilon_x}$, and the sum of $\sigma^2_\mu + \sigma^2_{\epsilon_y}$ would be identified. The classic psychometric assumption of equal reliability means that the error variances are proportional to the true variances, e.g., if the reliability of x and y were equal then $\lambda = \sigma^2_{\epsilon_y} / \sigma^2_{\epsilon_x} = \hat{\sigma}^2_y / \hat{\sigma}^2_x$. This equal reliability assumption in combination with the congeneric measure of X would identify σ^2_μ and $\sigma^2_{\epsilon_y}$ separately. In principle, this congeneric measure serves much the same purpose as the econometrician's use of an "instrumental variable" (Johnston, 1963, sec. 6.5) i.e., a variable which is independent of the measurement errors ϵ_x and ϵ_y. For example, if an instrumental variable z were available for Isaac's model, the observed vector would be (y, x, z), the factors are (X, Z, ϵ_y, ϵ_x), B = an identity matrix, Ψ and $\Theta = 0$,

$$\Lambda = \begin{bmatrix} \beta & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

$$\text{and} \quad \Phi = \begin{bmatrix} \sigma^2_X & C_{XZ} & 0 & 0 \\ C_{XZ} & \sigma^2_Z & 0 & 0 \\ 0 & 0 & \sigma^2_{\epsilon_y} & 0 \\ 0 & 0 & 0 & \sigma^2_{\epsilon_x} \end{bmatrix}$$

For convenience a factor Z has been defined identical to z (i.e., $\sigma_Z = \sigma_z$), however we could have considered the model $z = Z + \epsilon_z$ which would have identified the parameters $\beta, \sigma^2_X, \sigma^2_{\epsilon_y}, \sigma^2_{\epsilon_x}$, and C_{XZ} but not σ^2_Z and $\sigma^2_{\epsilon_z}$. Jöreskog's general model thus allows the analyst considerable flexibility in his choice of econometric and/or psychometric procedures for dealing with errors of measurement.

In summary, we recommend use of Jöreskog's general model because: (a) It is unnecessary to have estimating formulae for each special case, especially since such formulae do not apply to overidentified models. (b) Attention is focussed on the problem of identification which is prerequisite to any understanding of the results. (c) Given multivariate normality of observed variables, a chi-squared goodness of fit test is available. If for example, in Isaac's case #4 we wished to test the hypothesis that β was a given value, then the increase in χ^2 (with one degree of freedom) resulting from changing β to a fixed parameter, is a test of the tenability of this hypothesis. (d) A variety of assumptions may be used singly or in combination, so that whatever information is available may be incorporated, hopefully achieving an overidentified model which can be tested for fit.

Footnote

[1] The estimating formula for $\hat{\beta}$ in case #3, given by equation (3) in Isaac (1970, pg. 215) has a λ left out of the denominator. Kendall and Stuart (1961) recommend that the positive root be used. Johnston (1963, pg. 154), however, recommends that the positive root should be used when $cov\ (x,y)$ is positive and the negative root when $cov(x,y)$ is negative.

[†] The research reported herein was performed pursuant to Grant No. OEG-2-700033(509) with the United States Department of Health, Education, and Welfare and the Office of Education.

Bibliography

Isaac, P. Linear regression, structural relations, and measurement error. *Psychological Bulletin*, 1970, *74*, 213-218.

Johnston, J. *Econometric methods*. New York: McGraw Hill, 1963.

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

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

Kendall, M.G., & Stuart, A. *The advanced theory of statistics*. Vol. II Inference and relationship. London: Charles Griffin & Co., 1961.

12/21/70
rbh

RB-71-22

RESEARCH BULLETIN

A CONGENERIC MODEL FOR PLATONIC TRUE SCORES

Charles E. Werts, Robert L. Linn
and Karl Jöreskog

Educational Testing Service
Princeton, New Jersey
May 1971

A CONGENERIC MODEL FOR PLATONIC TRUE SCORES

Charles E. Werts, Robert L. Linn, and Karl Jöreskog

Abstract

To resolve a recent controversy between Klein and Cleary and Levy, a model for dichotomous congeneric items is presented which has mean errors of zero, dichotomous true scores that are uncorrelated with errors, and errors that are mutually uncorrelated.

A CONGENERIC MODEL FOR PLATONIC TRUE SCORES[1]

Charles E. Werts, Robert L. Linn, and Karl Jöreskog

In a discussion of platonic true scores, Klein and Cleary (1967) state that the use of platonic true scores makes the assumptions of classical test theory generally untenable. They illustrate their argument with dichotomous items and a dichotomous true score and show that: "The classical test theory formulation $\sigma_X^2 = \sigma_T^2 + \sigma_E^2$, can only be true if the mean error is not zero" (Klein & Cleary, 1967, p. 78). This statement is based on the following definitions of observed (X), true (T), and error (E) scores:

$$T = \left\{\begin{matrix} \text{1 if phenomenon is present} \\ \text{0 if phenomenon is absent} \end{matrix}\right\},$$

$$X = \left\{\begin{matrix} \text{1 if phenomenon is rated as present} \\ \text{0 if phenomenon is rated as absent} \end{matrix}\right\},$$

and $E = X - T$. Klein and Cleary go on to consider two parallel dichotomous items, X_1 and X_2, and show that the covariance between E_1 and E_2 is positive when the errors, E_1 and E_2, have zero means. With correlated error scores, the correlation between two parallel items overestimates the item reliabilities. In response, Levy (1969) argued that the classical assumptions can be shown to hold for a dichotomous item if

$$X = \left\{\begin{matrix} \text{a if phenomenon is rated as present} \\ \text{b if phenomenon is rated as absent} \end{matrix}\right\},$$

true scores (T) are defined as above and $E = X - T$ as before. This modification will indeed make it possible for the mean error to be zero and the covariance between T and E to be zero. As Klein and Cleary (1969) note, however, Levy does not provide a means of solving for "a" and "b" without knowledge of T . In any practical application, T would be unknown and therefore "a" and "b" would be unknown. Also, no way of obtaining item reliabilities is presented. The purpose of this paper is to provide an alternative formulation which allows for the model parameters to be determined given the structural specification of zero mean error and no correlation among errors for different items or between errors and true scores. Our approach is drawn from latent structure analysis (Anderson, 1959) for the special case of dichotomous latent variables.

I. A Congeneric Model for Dichotomous Items

The equation for congeneric tests is given by Jöreskog (1968, 1970, 1971) as

$$X_{ij} = B_{jT}T_i + I_j + E_{ij} \quad , \qquad (1)$$

where T_i is the true score for person i ,

X_{ij} is the observed score on item j for person i ,

B_{jT} is the slope of the X_{ij} on T_i regression line,

I_j is the intercept of this regression line, and

E_{ij} is the error for person i on item j .

To illustrate the application of this definition to the case in which X_{ij} and T_i are both dichotomous (scored 1, 0), consider the case of three items, which is the minimum number of items required to identify model parameters uniquely, given experimentally independent measures. The equations are

$$X_1 = B_{1T}T + I_1 + E_1 \quad , \qquad (1a)$$

$$X_2 = B_{2T}T + I_2 + E_2 \quad , \qquad (1b)$$

$$X_3 = B_{3T}T + I_3 + E_3 \quad , \qquad (1c)$$

where the E 's are mutually uncorrelated and are uncorrelated with T .[2] In the case of dichotomous variables

$$B_{jT} = \frac{P\{X_j = 1, T = 1\} - P\{X_j = 1\}P\{T = 1\}}{P\{T = 1\}P\{T = 0\}} = P\{X_j = 1|T = 1\} - P\{X_j = 1|T = 0\}$$

and

$$I_j = P\{X_j = 1\} - B_{jT}P\{T = 1\} = P\{X_j = 1|T = 0\} \ .$$

This model is somewhat more complicated than the model considered by Klein and Cleary (1967) where $X = T + E$ with X , T , and E all taking values of 0 or 1. In essence, the congeneric model is equivalent to the model suggested by Levy (1969) if his "a" and "b" are allowed to vary from item to item. For a given item, "a_j" would equal $(1 - I_j)/B_{jT}$, "b_j" would equal $-I_j/B_{jT}$, and Levy's error would equal the error of equations 1, 2, or 3 divided by B_{jT} . To illustrate the point that the congeneric model does allow for the traditional psychometric assumptions in the dichotomous case; consider the following example constructed using the equations provided by Anderson (1959, sec. 2.4).

1. The θ_j (proportion of false negatives, i.e., $P\{X_j = 0|T = 1\} = P\{X_j = 0,T = 1\} \div P_T$), ϕ_j (proportion of false positives, i.e., $P\{X_j = 1|T = 0\} = P\{X_j = 1,T = 0\} \div (1 - P_T)$) , and P_T (the true proportion $P\{T = 1\}$) are given as:

$$\theta_1 = .30 \ , \ \theta_2 = .40 \ , \ \theta_3 = .10 \ ;$$
$$\phi_1 = .10 \ , \ \phi_2 = .50 \ , \ \phi_3 = .30 \ ;$$
$$P_T = .60 \ , \text{ and } Q_T = 1 - P_T = .40 \ .$$

2. The expected marginal distributions $(P_j = \text{Prob } \{X_j = 1\})$ are $P_j = (1 - \theta_j)P_T + \phi_j Q_T$, i.e., $P_1 = .46$, $P_2 = .56$, and $P_3 = .66$.

3. The expected joint probabilities for pairs of items, $P_{jj'}$ = Prob $\{X_j = 1, X_{j'} = 1\} = (1 - \theta_j)(1 - \theta_{j'})P_T + \phi_j\phi_{j'}Q_T$ $(j \neq j')$ are: $P_{12} = .272$, $P_{13} = .390$, and $P_{23} = .384$.

4. The expected joint probability for three items, $P_{jj'j''}$ = Prob $\{X_j = 1, X_{j'} = 1, X_{j''} = 1\} = (1 - \theta_j)(1 - \theta_{j'})(1 - \theta_{j''})P_T + \phi_j\phi_{j'}\phi_{j''}Q_T$ $(j \neq j' \neq j'')$ is $P_{123} = .2328$.

5. The regression weights $(B_{jt} = 1 - \theta_j - \phi_j)$ are $B_{1T} = .60$, $B_{2T} = .10$, and $B_{3T} = .60$.

6. The intercepts $(I_j = P_j - B_{jT}P_T = \phi_j)$ are $I_1 = .10$, $I_2 = .50$, and $I_3 = .30$. The possible events for combinations of the three items and the proportion of people in each event are shown in Table 1. The means of the errors are zero, the true score is uncorrelated with the errors and the errors are uncorrelated with each other.

Insert Table 1 about here

II. Identification

In an actual problem the situation would be reversed from the example shown in section I, i.e., the probabilities P_1, P_2, P_3, P_{12}, P_{13}, P_{23}, and P_{123} correspond to observed scores, and it would be desirable to identify the seven parameters, θ_1, θ_2, θ_3, ϕ_1, ϕ_2, ϕ_3, and P_T . In principle, one could solve the seven equations for this purpose:

$$P_1 = (1 - \theta_1)P_T + \phi_1Q_T \ , \tag{2a}$$

$$P_2 = (1 - \theta_2)P_T + \phi_2Q_T \ , \tag{2b}$$

$$P_3 = (1 - \theta_3)P_T + \phi_3Q_T \ , \tag{2c}$$

$$P_{12} = (1 - \theta_1)(1 - \theta_2)P_T + \phi_1\phi_2Q_T \ , \qquad (2d)$$

$$P_{13} = (1 - \theta_1)(1 - \theta_3)P_T + \phi_1\phi_3Q_T \ , \qquad (2e)$$

$$P_{23} = (1 - \theta_2)(1 - \theta_3)P_T + \phi_2\phi_3Q_T \ , \qquad (2f)$$

$$P_{123} = (1 - \theta_1)(1 - \theta_2)(1 - \theta_3)P_T + \phi_1\phi_2\phi_3Q_T \ . \qquad (2g)$$

The solution to these equations is facilitated by noting that in the congeneric model the expected covariance $(C_{jj'})$ between two items is given by

$$C_{jj'} = B_{jT}B_{j'T}V_T \ ,$$

where V_T is the variance of T . Translating into probabilities:

$$(P_{jj'} - P_jP_{j'}) = (1 - \theta_j - \phi_j)(1 - \theta_{j'} - \phi_{j'})P_TQ_T \ . \qquad (3)$$

This means that

$$C_{12} = P_{12} - P_1P_2 = (1 - \theta_1 - \phi_1)(1 - \theta_2 - \phi_2)P_TQ_T \ , \qquad (4a)$$

$$C_{13} = P_{13} - P_1P_3 = (1 - \theta_1 - \phi_1)(1 - \theta_3 - \phi_3)P_TQ_T \ , \qquad (4b)$$

$$C_{23} = P_{23} - P_2P_3 = (1 - \theta_2 - \phi_2)(1 - \theta_3 - \phi_3)P_TQ_T \ . \qquad (4c)$$

These equations can be solved for

$$(1 - \theta_1 - \phi_1)^2 P_TQ_T = \frac{C_{12}C_{13}}{C_{23}} = B_{1T}^2P_TQ_T \ , \qquad (5a)$$

$$(1 - \theta_2 - \phi_2)^2 P_TQ_T = \frac{C_{12}C_{23}}{C_{13}} = B_{2T}^2P_TQ_T \ , \qquad (5b)$$

$$(1 - \theta_3 - \phi_3)^2 P_TQ_T = \frac{C_{13}C_{23}}{C_{12}} = B_{3T}^2P_TQ_T \ . \qquad (5c)$$

The triple covariance C_{123} is defined (Boudon, 1968, p. 226) as the expectation of the products of the deviations of all three variables simultaneously, which is equal in the dichotomous case to $C_{123} = P_{123} - P_1(P_{23} - P_2P_3) - P_2(P_{13} - P_1P_3) - P_3(P_{12} - P_1P_2) - P_1P_2P_3$. (6)

Using equations (2a, b, c, d, e, f) equation (6) may be translated to $C_{123} = B_{1T}B_{2T}B_{3T}P_TQ_T(Q_T^2 - P_T^2)$ and from equations (5a, b, c) we obtain

$$C_{123} = \frac{\sqrt[3]{C_{12}C_{13}C_{23}}(Q_T^2 - P_T^2)}{\sqrt{P_TQ_T}} \quad . \qquad (7)$$

Applying these equations to our example,

1. Compute covariances by equations (4a, b, c):

   $C_{12} = .0144$, $C_{13} = .0864$, and $C_{23} = .0144$.

2. Using equation (6) compute $C_{123} = -.001728$.

3. From equation (7),

$$\frac{C_{123}}{\sqrt{C_{12}C_{13}C_{23}}} = -.4082 = \frac{(Q_T^2 - P_T^2)}{\sqrt{P_TQ_T}} \quad .$$

4. Solving for $P_T = 1 - Q_T$ we obtain $P_T = .60$.

5. From equations (5a, b, c) and substituting in this value of P_T ,

   $B_{1T} = .60$,

   $B_{2T} = .10$,

   $B_{3T} = .60$.

6. It can be shown (equations 2a, b, c) that $\phi_j = P_j - B_{jT}P_T$ permitting calculation of $\phi_j = I_j$:

$$I_1 = \phi_1 = .10 \quad ,$$

$$I_2 = \phi_2 = .50 \quad ,$$

$$I_3 = \phi_3 = .30 \quad .$$

7. Since $\theta_j = 1 - B_{jT} - \phi_j$,

$$\theta_1 = .30 \quad ,$$

$$\theta_2 = .40 \quad ,$$

$$\theta_3 = .10 \quad .$$

8. Item reliabilities R_{jj} are $R_{jj} = B^2_{jT}P_TQ_T/P_jQ_j$, i.e.,

$$R_{11} = .3478 \quad ,$$

$$R_{22} = .0097 \quad ,$$

$$R_{33} = .3850 \quad .$$

In the case of three congeneric items the model parameters are just identified, i.e., there are seven equations in seven unknowns, which is the reason that the parameters may be obtained as an exact function of the observed probabilities. In the case of overidentified models one of the estimating procedures discussed by Anderson (1959) can be used. One procedure minimizes a χ^2 function of the observed probabilities (P_O) and the expected probabilities (P_E) generated as a function of the parameter estimates (Cochran, 1968; Mote & Anderson, 1965). In the general case of J items there will be $(2^J - 1)$ independent observed probabilities in the cross-tabulation table from which $(2J + 1)$ parameters are to be estimated. In the special case of two items of equal accuracy the reliability is the correlation between these items, but the model parameters cannot be identified

(Cochran, 1968, sec. 6) since $P_E\{X_1 = 1, X_2 = 0\} = P_E\{X_1 = 0, X_2 = 1\}$, i.e., there are only two independent probabilities to estimate three parameters (θ, ϕ, P_T).

III. Variations

It is sometimes the case that three items with errors that are uncorrelated with true scores or errors of other items are available but one of these measures another variable, i.e.,

$$X_1 = B_1T_1 + I_1 + E_1 \;,$$
$$X_2 = B_2T_1 + I_2 + E_2 \;, \qquad (8)$$
$$X_3 = B_3T_2 + I_3 + E_3 \;.$$

In econometrics X_3 is called an "instrumental" variable (Johnston, 1963, p. 165). The equation for X_3 can be transformed into

$$X_3 = B_3^*T_1 + I_3^* + E_3^* \;, \qquad (8a)$$

where

$$B_3^* = B_{T_2T_1}B_3 \;.$$

B_3^* is identified but $B_{T_2T_1}$ and B_3 are not. In the case of dichotomous variables, therefore, the true proportion P_{T_1} may be estimated as shown in section II by treating X_3 as a congeneric measure of T_1 and $B_3^* = (1 - \theta_3 - \phi_3)(1 - \theta_{T_1} - \phi_{T_1})$, where $\theta_{T_1} = P\{T_2 = 0|T_1 = 1\}$ and $\phi_{T_1} = P\{T_2 = 1|T_1 = 0\}$. The validity of such an analysis is dependent on the correctness of the independence assumption.

The above analysis can be extended to the case of four items with mutually uncorrelated errors and no correlation between error and true scores, two of each measuring different variables:

$$\begin{aligned} X_1 &= B_1T_1 + I_1 + E_1 \ , \\ X_2 &= B_2T_1 + I_2 + E_2 \ , \\ X_3 &= B_3T_2 + I_3 + E_3 \ , \\ X_4 &= B_4T_2 + I_3 + E_3 \ . \end{aligned} \tag{9}$$

Following the above line of reasoning all parameters in this model (P_{T_1}, $P\{T_1 = 1, T_2 = 1\}$, P_{T_2}, θ_1, θ_2, θ_3, θ_4, ϕ_1, ϕ_2, ϕ_3, and ϕ_4) may be identified. There are 15 independent proportions in the cross-tabulation table, so that the minimized χ^2 would have four degrees of freedom. In principle, a measure of the tenability of certain assumptions is obtained from changes in the χ^2. For example, if it were desired to test the hypothesis that X_1 and X_2 were of equal accuracy, increases in the total χ^2 (with two degrees of freedom), resulting from setting $\theta_1 = \theta_2$ and $\phi_1 = \phi_2$, would be an indicator of the tenability of this hypothesis.

References

Anderson, T. W. Some scaling models and estimation procedures in the latent class model. In U. Grenander (Ed.), Probability and statistics, The Harold Cramér volume. New York: Wiley, 1959. Pp. 9-38.

Boudon, R. A new look at correlation analysis. In H. M. Blalock & A. B. Blalock (Eds.), Methodology in social research. New York: McGraw-Hill, 1968. Pp. 199-235.

Cochran, W. G. Errors of measurement in statistics. Technometrics, 1968, 10, 637-666.

Johnston, J. Econometric methods. New York: McGraw-Hill, 1963.

Jöreskog, K. G. Statistical models for congeneric test scores. Proceedings of the 76th Annual Convention of the American Psychological Association, 1968, 213-214.

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

Jöreskog, K. G. Statistical analysis of sets of congeneric tests. Psychometrika, 1971, 36, in press.

Klein, D. F., & Cleary, T. A. Platonic true scores and error in psychiatric rating scales. Psychological Bulletin, 1967, 68, 77-80.

Klein, D. F., & Cleary, T. A. Platonic true scores: Further comment. Psychological Bulletin, 1969, 71, 278-280.

Levy, P. Platonic true scores and rating scales: A case of uncorrelated definitions. Psychological Bulletin, 1969, 71, 276-277.

Mote, V. L., & Anderson, R. L. An investigation of the effect of misclassification on the properties of χ^2-tests in the analysis of categorical data. Biometrika, 1965, 52, 95-109.

Footnotes

[1] The research reported herein was performed pursuant to Grant No. OEG-2-700033(509) with the United States Department of Health, Education, and Welfare and the Office of Education.

[2] The true scores are not independent of the error scores or errors of each other, as is assumed in Anderson's (1959) derivations; however, for our purposes the assumption that these variables are uncorrelated yields the same formulas.

Table 1

Possible Events for Three Congeneric Dichotomous Items

Proportion of People	T	X_1	X_2	X_3	E_1	E_2	E_3
.2268	1	1	1	1	.3	.4	.1
.0252	1	1	1	0	.3	.4	-.9
.1512	1	1	0	1	.3	-.6	.1
.0168	1	1	0	0	.3	-.6	-.9
.0972	1	0	1	1	-.7	.4	.1
.0108	1	0	1	0	-.7	.4	-.9
.0648	1	0	0	1	-.7	-.6	.1
.0072	1	0	0	0	-.7	-.6	-.9
.0060	0	1	1	1	.9	.5	.7
.0140	0	1	1	0	.9	.5	-.3
.0060	0	1	0	1	.9	-.5	.7
.0140	0	1	0	0	.9	-.5	-.3
.0540	0	0	1	1	-.1	.5	.7
.1260	0	0	1	0	-.1	.5	-.3
.0540	0	0	0	1	-.1	-.5	.7
.1260	0	0	0	0	-.1	-.5	-.3

Estimating True Scores and True Group Means From Multiple Independent Measures

Charles E. Werts and Robert L. Linn

Abstract

Given multiple independent measures of an underlying true factor and information on group membership it is possible to compute a set of observed group means for each measure. Given a least three tests, these sets of means may be used to compute the reliability of the means for each test. The procedure for estimating true scores from the reliabilities of the individual tests and the group means is derived.

Estimating True Scores and True Group Means From Multiple Independent Measures[1]

Charles E. Werts and Robert L. Linn

The classical approach to estimating true scores given group membership information is to use the formula:

$$\hat{T}_{ij} = \bar{X}_j + R_{xx}(X_{ij} - \bar{X}_j) \qquad (1)$$

where $\hat{T}_{ij}$ is the estimated true score,

$\bar{Y}_j$ is the observed mean of group j ,

R_{xx} is the test reliability, assumed homogeneous across groups,

X_{ij} is the observed score for person i in group j .

If two parallel tests were available the reliability could be computed as the correlation between tests, however, two sets of observed individual values and group means would be observed. The estimation problem is to use both sets of data to obtain a better true score estimate than could be obtained from either. The general problem of using group information to estimate true scores given multiple measures will be considered in this paper.

For illustrative purposes consider the case where congeneric measures of an underlying true score factor are available. Congeneric measures (X_{ijk}) are related to the true score (T_{ij}):

[1] The research reported herein was performed pursuant to Grant No. OEG-1-6-061830-0650 Project No. 6-1830 with the Office of Education, U. S. Department of Health, Education and Welfare.

$$X_{ijk} = B_k T_{ij} + M_k + e_{ijk} \quad ,$$

where X_{ijk} is the observ. l value on person i in group j for test k ,

T_{ij} is the underlying true factor, (2)

B_k is the slope of the k^{th} test on T_{ij} ,

M_k is the intercept of the regression line of the k^{th} test scores on the true score,

e_{ijk} are error components for individual i on test k with zero mean for all levels of T_{ij} .

Equation (2) shows that congeneric tests may differ in units of measurement, reliability and mean, but that they all load on the same underlying factor. Three tests is the minimum number needed to solve for the reliability of each test (Lord & Novick, 1968, equation 9.12.4). The group means may be obtained for each test and from equation (2) it follows that:

$$\bar{X}_{jk} = B_k \bar{T}_j + M_k + \bar{e}_{jk} \quad , \qquad (3)$$

where $\bar{X}_{jk}$ is the observed group mean for group j on test k and $\bar{T}_j$ is the group mean on the true score.

For a given test it is useful to derive the condition under which the observed group means do not help to estimate the true score. In the prediction equation for the true score, $T_{ij} = B'X_{ijk} + B''\bar{X}_{jk} + e'_{ijk}$, the condition that the group means do not help is that $B'' = 0$. By definition:

$$B'' = \frac{V_{X_k} C_{T \bar{X}_k} - C_{T X_k} C_{X_k \bar{X}_k}}{V_{X_k} V^2_{\bar{X}_k} - C_{X_k \bar{X}_k}} \quad ,$$

where V_{X_k} is the variance of X_{ijk} ,

$V_{\bar{X}_k}$ is the variance of $\bar{X}_{jk}$,

$C_{T\bar{X}_k}$ is the covariance of T_{ij} and $\bar{X}_{jk}$,

C_{TX_k} is the covariance of T_{ij} and X_{ijk} , and

$C_{X_k\bar{X}_k}$ is the covariance ~~of X_{ijk}~~ and X_{ijk} .

It follows that B" = 0 implies $V_{X_k} \; C_{T\bar{X}_k} = C_{TX_k} \; C_{X_k\bar{X}_k}$. Since it can be shown that $C_{T\bar{X}_k} = C_{\bar{T}\bar{X}_k}$ and $C_{X_k\bar{X}_k} = V_{\bar{X}_k}$, B" = 0 means that $C_{\bar{T}\bar{X}_k}/V_{\bar{X}_k} = C_{TX_k}/V_{X_k}$ or $B_{\bar{T}\bar{X}_k} = B_{TX_k}$. We know, however, that $B_{\bar{X}_k\bar{T}} = B_{X_kT} = B_k$ therefore:

$B_{\bar{T}\bar{X}_k} \; B_{\bar{X}_k\bar{T}} = B_{TX_k} \; B_{X_kT}$ or $R_{\bar{X}_k\bar{T}} = R_{X_kT}$, where $R^2_{X_kT}$ is the reliability of test k .

In other words for a given test, the observed group means will not improve the prediction of the true score when the reliability of the means equals the reliability of the individual scores for each test. Since it is generally found that group means have a higher reliability than the individual scores, knowledge of group means can usually be expected to improve the estimation of true scores.

Our general strategy for estimating the true score T_{ij} will be to derive expressions for the correlation of the true score with each set of the observed individual test scores and of each set of observed group means. These correlations and the set of correlations among the observed variables (X_{ijk} and $\bar{X}_{jk}$) then permits us to solve for the standardized partial regression weights for predicting T_{ij} from the observed variables. The correlation of $\bar{X}_{jk}$ with T_{ij} ($R_{\bar{X}_kT}$) can be derived:

a. From equations (2) and (3) it follows that

$$C_{T\bar{\bar{X}}_k} = B_k \, C_{\bar{\bar{T}}T} \,,$$

where $C_{\bar{\bar{T}}T}$ is the covariance of $T_{\bar{j}}$ and T_{ij}.

b. Since $C_{\bar{\bar{T}}T} = V_{\bar{\bar{T}}}$ (the weighted variance of the means)

$$C_{T\bar{\bar{X}}_k} = B_k \, V_{\bar{\bar{T}}} \quad \text{and}$$

$$R_{T\bar{\bar{X}}_k} = \frac{B_k \, V_{\bar{\bar{T}}}}{\sqrt{V_T \, V_{\bar{\bar{X}}_k}}} = R_{\bar{\bar{T}}\bar{\bar{X}}_k} \sqrt{\frac{V_{\bar{\bar{T}}}}{V_T}} \,.$$

c. By definition $B_k = R_{TX_k} \sqrt{V_{X_k} \div V_T} = R_{\bar{\bar{T}}\bar{\bar{X}}_k} \sqrt{V_{\bar{\bar{X}}_k} \div V_{\bar{\bar{T}}}}$

therefore: $$\sqrt{\frac{V_{\bar{\bar{T}}}}{V_T}} = \frac{R_{\bar{\bar{T}}\bar{\bar{X}}_k}}{R_{TX_k}} \cdot \sqrt{\frac{V_{\bar{\bar{X}}_k}}{V_{X_k}}} \,.$$

d. By substitution

$$R_{T\bar{\bar{X}}_k} = \frac{(R_{\bar{\bar{T}}\bar{\bar{X}}_k})^2}{R_{TX_k}} \cdot \sqrt{\frac{V_{\bar{\bar{X}}_k}}{V_{X_k}}} \,. \qquad (4)$$

Since the standard deviations of the means ($\sqrt{V_{\bar{\bar{X}}_k}}$) and of the individual values ($\sqrt{V_{X_k}}$) can be computed directly from the data, the correlation of the observed group means for test k with the true score can be computed from the reliabilities for the means ($R_{\bar{\bar{T}}\bar{\bar{X}}_k}{}^2$) and the individual scores ($R^2_{TX_k}$).

Since equation (2) is a factor analytic model with one common factor T_{ij} and equation (3) a factor model with the common factor $\bar{T}_j$, the reliabilities correspond to the square of the corresponding (standardized) factor loading. Jöreskog (1969b) discusses the factor analysis of congeneric measures in considerable detail. With more than three measures the model can be tested to see how consistent the congeneric assumption is with the data. Stronger assumptions about the tests (e.g., equivalency) can be readily incorporated into the analysis.

In summary then, the computational procedure involves:

1. Calculation of the group means for each of the k tests,

2. Creation of a new set of k variables by assigning to each individual the mean of his group on each of the k tests,

3. Intercorrelation of all 2 k variables and computation of standard deviations.

4. Factor analysis of the k sets of test scores using as input the correlations among those tests from step 3. If Jöreskog's (1969a) confirmatory factor analysis procedure is used for this purpose a chi-squared goodness of fit measure will be obtained along with maximum likelihood estimates of the factor loadings (which are squared to obtain reliability estimates for that test).

5. Factor analysis of the k sets of group means using as input the correlations among those tests from step 3. If desired, factor score estimates of the true group means may be obtained.

6. The correlations of the k sets of group means with the true score can be computed from equation (4) where R_{TX_k} is the factor loading for test k computed in step 4, $R_{\bar{T}\bar{X}_k}$ is the factor loading for test k group means computed in step 5, $\sqrt{V_{X_k}}$ is the standard deviation of the individual scores

on test k computed in step (3) and $\sqrt{V_{\bar{X}_k}}$ is the standard deviation of the group means on test k computed in step (3).

7. The next step is computation of the standardized regression weights for predicting the true score from the 2 k observed variables. The correlations among the observed variables from step (3), the correlations of the k sets of test scores with the true score are the factor loadings from step (4), and the correlations of the k sets of group means from step (6) may be used in the "normal equations" to solve for the desired regression weights (Walker & Lev, 1953, pgs. 324-336). These weights could in turn be used to estimate a standardized true score for each individual from his observed test scores and group mean on each of the tests.

Variations

The above procedure requires that the means for each group on each test be computed and that these mean values be assigned to individuals so that a set of variables is created which may be intercorrelated. The advantage of this approach is that the reliabilities of the means may be computed for each test and the true group means estimated as factor scores. Instead of this analysis a factor analytic model might be postulated to account for all the correlations among the 2 k observed (X_{ijk} and $\bar{X}_{jk}$) variables. This model would have:

1. A total of (2k + 2) factors including T_{ij} , $\bar{T}_j$, and a residual factor for each of the observed 2k variables.

2. All residual factors involving different tests would be assumed independent corresponding to the congeneric assumption, whereas each pair of residuals corresponding to the same test data would be nonindependent (because the group means are computed from the individual scores for a given test).

3. Each of the X_{ijk} would load on T_{ij} and each of the $\bar{X}_{jk}$ would load on $\bar{T}_j$. T_{ij} and $\bar{T}_j$ would be nonindependent, their correlation being equal to the true correlation ratio.

4. Because reliabilities are desired the correlation matrix would be the basic input data and the variance of T_{ij} and $\bar{T}_j$ would be set equal to unity.

5. This factor model would have a vector of order 2 k of observed scores and a vector of order 2 k + 2 factors and no vector of unique scores. When $k > 3$ the model is overidentified and Jöreskog's (1969c) confirmatory factor analysis program could be used for estimation purposes. The program would estimate the factor loadings, the error variances, and error covariances among nonindependent residuals.

6. If the analysis were repeated specifying that for each test the loading of X_{ijk} on T_{ij} were equal to the loading of $\bar{X}_{jk}$ on $\bar{T}_j$, then a test of the assumption that for each test the reliability of the means equalled that of the individual scores would be the change in the chi-squared with k degrees of freedom.

In the event that it is desired only to improve the estimation of the overall true scores using group information a more direct approach may be taken by coding the group information as a set of dummy variables (Z_j). The model for this analysis given three congeneric tests is depicted in Figure 2:

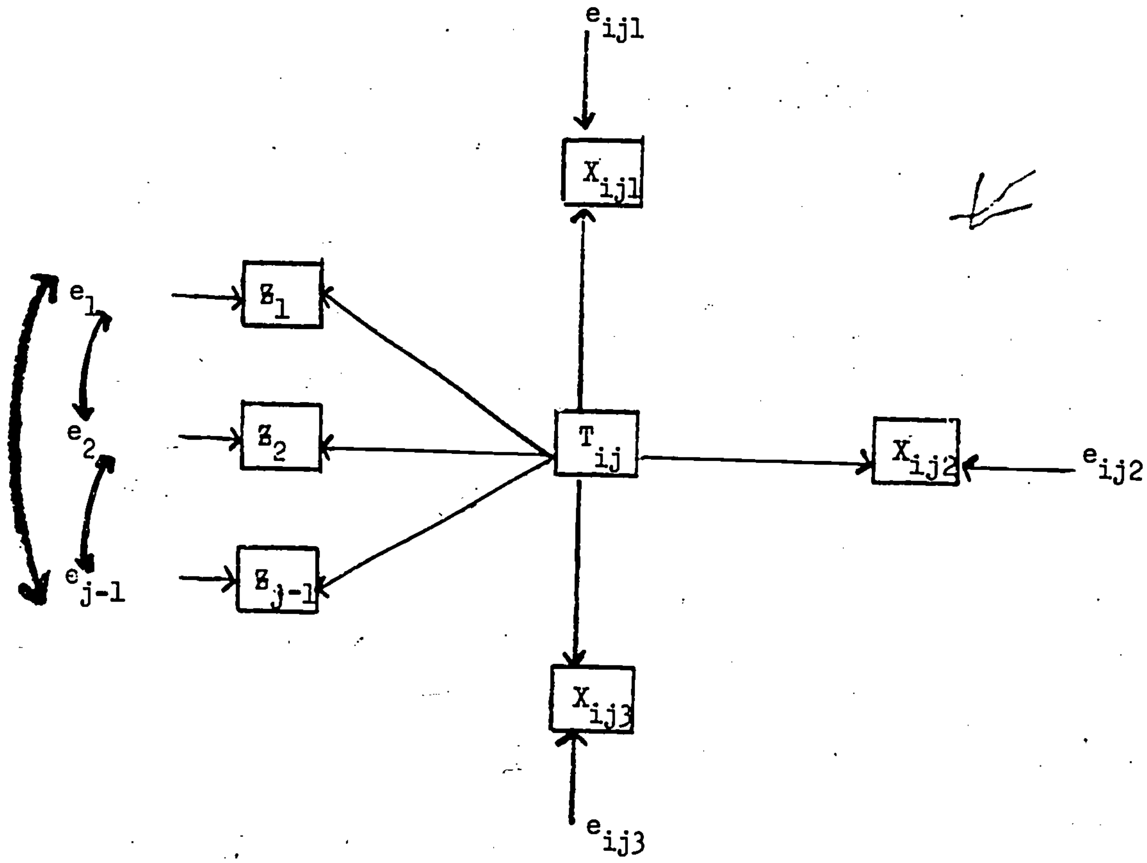


Figure 2. Estimating true scores with dummy variables.

The covariance between the dummy variables and an observed set of test scores will be a function of true mean differences between groups and the reliability of the means for that test. In essence, the dummy variables add information about the true group means to the estimation of T_{ij} . Since the last dummy variable is perfectly predictable from the other dummy variables it may be deleted from the computations. Since the dummy variables in part represent overlapping information about group membership (e.g., a person in one group is not in the next group) the residuals are shown as correlated in Figure 2. The factors are now (T_{ij} , e_{ij1} , e_{ij2} , e_{ij3} , e_{z1} , e_{z2} , ... , $e_{z(j-1)}$ and the observed vector (X_{ij1} , X_{ij2} , X_{ij3} , Z_1 , Z_2 ,..., $Z_{(j-1)}$. The hypothesized factor loading matrix is:

$$\Lambda = \left[\begin{array}{llllll}
B_1 & 1 & 0 & 0 & 0 & 0 \ldots 0 \\
B_2 & 0 & 1 & 0 & 0 & 0 \ldots 0 \\
B_3 & 0 & 0 & 1 & 0 & 0 \ldots 0 \\
B_{z1} & 0 & 0 & 0 & 1 & 0 \ldots 0 \\
B_{z2} & 0 & 0 & 0 & 0 & 1 \ldots 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \ddots \vdots \\
B_{z(j-1)} & 0 & 0 & 0 & 0 & 0 \ldots 1
\end{array}\right]$$

The hypothesized variance-covariance matrix of the factors is:

$$\Phi = \left[\begin{array}{lllllll}
1 & & & & & & \\
0 & V_{e_{ij1}} & & & & \text{Symmetric} & \\
0 & 0 & V_{e_{ij2}} & & & & \\
0 & 0 & 0 & V_{e_{ij3}} & & & \\
0 & 0 & 0 & 0 & V_{e1} & & \\
0 & 0 & 0 & 0 & C_{e_1e_2} & V_{e_2} & \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\
0 & 0 & 0 & 0 & C_{e_1e_{(j-1)}} & C_{e_2e_{(j-1)}} & \ldots V_{e_{(j-1)}}
\end{array}\right]$$

This approach becomes computationally awkward when the number of groups becomes large, in which case it may prove easier to first compute the means and proceed as shown in the previous section.

In passing, the relationships with a one way analysis of variance with a fallible dependent variable might be noted. The problem in that case would be whether the true means differed from one treatment group to the next, i.e., whether $V_{\bar{T}_j} > 0$. In the model used above to test for equal reliabilities this would correspond to the hypothesis that $\rho_{T\bar{T}} = 0$ since this correlation is the correlation ratio, i.e., $\rho_{T\bar{T}} = \sqrt{V_{\bar{T}} \div V_T}$. To test the hypothesis the analysis coula be rerun with $\rho_{T\bar{T}}$ a "fixed" parameter set = 0 and the difference in chi squared values (one degree of freedom) would be the appropriate significance test. One might consider using the congeneric model for the analysis of variance where the treatment effects are measured in terms of several symptoms which presumably reflect some underlying process which is not directly measured. Providing that the errors of measurement between symptoms are independent and the symptoms are linearily related to the underlying process; the congeneric model might provide a more valid test of the hypothesis.

cvs
3/10/70

Bibliography

Jöreskog, K. G. A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, 1969a, *34*, 183-202.

Jöreskog, K. G. Factoring the multitest-multioccasion correlation matrix. Research Bulletin 69-92, Princeton, N.J.: Educational Testing Service, July, 1969b.

Jöreskog, K. G. Statistical analysis of sets of congeneric tests. Research Bulletin 69-97, Princeton, N.J.: Educational Testing Service, December, 1969c.

Lord, F. M., & Novick, M. R. *Statistical theories of mental test scores*. Reading, Mass.: Addison-Wesley, 1968.

Walker, H. M. & Lev, J. *Statistical inference*. New York: Holt, 1953.

ERRORS OF INFERENCE DUE TO ERRORS OF MEASUREMENT

Robert L. Linn and Charles E. Werts

Educational Testing Service

Abstract

Failure to consider errors of measurement when using partial correlation or analysis of covariance techniques can result in erroneous conclusions. Certain aspects of this problem are discussed and particular attention is given to issues raised in a recent article by Brewer, Campbell, and Crano.

ERRORS OF INFERENCE DUE TO ERRORS OF MEASUREMENT[1,2]

Robert L. Linn and Charles E. Werts

Educational Testing Service

Brewer, Campbell, and Crano (1970) have justifiably criticized the use of partial correlation procedures in hypothesis testing research where errors of measurement are not taken into consideration. Ignoring measurement errors is much more serious when dealing with partial correlations than when dealing with simple zero-order correlations. In the latter case we know that the effect of errors of measurement, that are mutually uncorrelated and uncorrelated with true scores, is to reduce the absolute value of the zero-order correlation between the fallible measures. As Lord (1963) has pointed out, however, we cannot ordinarily know the effect of such errors of measurement on a partial correlation. Errors of measurement can increase or decrease the magnitude of a partial correlation and may even result in a partial correlation of a different sign.

As an alternative, Brewer et al. (1970) have suggested that factor analytic techniques be used to test a single-factor model before drawing conclusions about the nature of underlying conceptual variables. The purpose of the present paper is to reconsider the issues raised by these authors and the reasoning that led to their conclusions. Attention also will be given to some related arguments that were made in a recent attack on some commonly used methods for the evaluation of compensatory educational programs (Campbell & Erlebacher, 1970). Our thesis is that the basic problem is a lack of relevant information--a problem that cannot be resolved by the choice of a statistical procedure.

Relationship between Factor and Partial Correlation Analyses

Ignoring errors of measurement, the relationship between the loadings on a single common factor and the partial correlations in the case of three variables is straightforward. The squared factor loadings on a single common factor can be expressed:

$$a_i^2 = \frac{\rho_{ij}\rho_{ik}}{\rho_{jk}} \qquad (1)$$

for $i,j,k = 1,2,3$; $i \neq j \neq k$, where a_i is the factor loading on the single common factor for variable i and the ρ 's are the intercorrelations among the variables, i,j,k . When $\rho_{jk} = 0$, a_i^2 is undefined. Assuming none of the three zero-order correlations equal zero, the squared factor loading can be written as a function of the partial correlation, $\rho_{jk.i}$:

$$a_i^2 = 1 - C\rho_{jk.i} \quad , \qquad (2)$$

where

$$C = \frac{\sqrt{1 - \rho_{ij}^2}\sqrt{1 - \rho_{ik}^2}}{\rho_{jk}} \quad .$$

Provided that C is positive, it may be seen from (2) that when $\rho_{jk.i} = 0$, $a_i^2 = 1.0$ and when $\rho_{jk.i} < 0$, $a_i^2 > 1.0$.

Frederic Lord (personal communication) suggested that the relationship between the factor and partial correlation analyses could be clarified by an example such as the one depicted in Figure 1. Given $\rho_{X_1X_2} = .50$, the possible values of $\rho_{X_1X_3}$ and $\rho_{X_2X_3}$ are contained in the ellipse in Figure 1.

Regions of the figure that contain negative partial correlations are indicated. Factor loadings are denoted by a_i and regions that contain imaginary loadings or squared loadings greater than 1.0 are indicated.

Insert Figure 1 about here

On line segment $\underline{ac}$ $\rho_{23.1} = 0$ and $a_1 = 1.0$, on line segment $\underline{bd}$ $\rho_{13.2} = 0$ and $a_2 = 1.0$, and on line segments $\underline{aeb}$ and $\underline{cfd}$ $\rho_{12.3} = 0$ and $a_3 = 1.0$. Imaginary values of the a 's occur when one of the three zero-order correlations is negative while the other two are positive.

Bias in Partial Correlation

Brewer et al. (1970) argue that errors of measurement introduce a systematic bias into partial correlations. More specifically, they state: ". . . the assumption is made that the variable being partialled out contains no unique components and is measured without error. Using partialling techniques when these assumptions are not met introduces systematic bias toward the unparsimonious conclusion that more conceptual factors are involved in a phenomenon than may actually be the case" (Brewer et al., 1970, pp. 1-2). Although it is true that this may be the effect of a violation of the assumption of an error free measure, the bias may be in the opposite direction. It is easy to construct an example where the direction of the bias is toward a more parsimonious conclusion that fewer conceptual factors are involved in a phenomenon than is actually the case. Suppose, for example, that three latent variables (T_1 , T_2 , and T_3) had the following intercorrelations in the population:

$$\rho_{T_1T_2} = .6 \quad ,$$
$$\rho_{T_1T_3} = .6 \quad ,$$
$$\text{and} \quad \rho_{T_2T_3} = .18 \quad .$$

The correlation between T_2 and T_3 with T_1 partialed out is $-.28125$ and the corresponding conclusion is that more than one conceptual variable is involved in this phenomenon. Suppose, however, that only a fallible measure of the first variable, say X_1, was available, where

$$X_1 = T_1 + E_1$$

and E_1 is uncorrelated with T_1, T_2, or T_3. Further, assume that the variance of X_1 is equal to twice the variance of T_1 (i.e., the reliability of X_1 is .50). Under these conditions the resulting intercorrelations among T_2, T_3, and X_1 would be:

$$\rho_{X_1T_2} = .6\sqrt{.5} \doteq .424 \quad ,$$
$$\rho_{X_1T_3} = .6\sqrt{.5} \doteq .424 \quad ,$$
$$\rho_{T_2T_3} = .18 \quad .$$

The correlation between T_2 and T_3 with X_1 partialed out would be 0.0 which would result in the more parsimonious, but erroneous conclusion that a second conceptual variable is not required. There is no intention to imply by this illustration that the bias of errors of measurement is typically, or even frequently, in the direction of producing a partial correlation that is closer to zero. Rather the point is that the direction of the bias cannot be determined without imposing additional assumptions (e.g., all reliabilities and all zero-order and partial correlations among true scores are nonnegative) and/or obtaining additional information such as the reliabilities of the measures. Given classical test theory assumptions, an estimate of the partial correlation among underlying true scores may be obtained by simply applying standard corrections for attenuation to the zero-order correlations. 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, further, the partial corrected for attenuation may be seriously affected by sampling errors. These obstacles can hardly justify the use of an uncorrected coefficient that may have the wrong sign, however" (Lord, 1963, p. 36).

The Single Factor Model vs. Partial Correlations

As noted above, Brewer et al. (1970) have suggested that a single-factor model be tested before conclusions are drawn about the nature of underlying conceptual variables from partial correlations. We shall argue that partial correlation analyses and factor analyses are based on different models and pose different questions. Knowing that a single factor can reproduce the intercorrelations among three observed fallible variables is not sufficient to draw conclusions about the partial correlations among the underlying conceptual variables or true scores that correspond to the observed scores.

Assuming that three infallible measures (T_1 , T_2 , and T_3) have a multivariate normal distribution, the partial correlation between T_2 and T_3 with T_1 partialed out has a very simple interpretation. It is equal to the zero-order correlation between T_2 and T_3 for any subpopulation defined by a particular value of T_1. Thus, it provides a means of investigating the relationship between T_2 and T_3 with T_1 held constant in the above sense. The question of whether or not T_2 and T_3 are related when T_1 is held constant is not the same as the question answered by a test for single factoredness for the observed scores. This is, in principle, acknowledged by Brewer et al. (1970) in footnote number 3 where they discuss an example in which the control variable (I.Q.) has a factor loading of .43. They conclude that "...if one has 'factored out' a variable upon which I.Q. loads only .43, one has not in any meaningful sense 'factored out I.Q.'" (Brewer et al., 1970, p. 7). They go on to indicate that they are working on a technique of "focused factoring," wherein the control variables are used to define the factor. Hopefully this procedure would exclude from the communality of a control variable only that variance that properly might be considered error variance.

If the observed variables (X_i) are related to their underlying true scores (T_i) by the model,

$$X_i = T_i + E_i , \qquad i = 1,2,3 ,$$

where the errors (E_i) are mutually uncorrelated and are uncorrelated with the true scores, then (1) may be expressed in terms of the correlations among the true scores, $\rho_{T_iT_j}$, and the reliabilities of the observed measures, ρ_{ii}, i.e., the variance of T_i divided by the variance of X_i. Thus

$$a_i^2 = \rho_{ii} \frac{\rho_{T_iT_j}\rho_{T_iT_k}}{\rho_{T_jT_k}} . \qquad (3)$$

The correlation between T_j and T_k with T_i partialed out is proportional to

$$\rho_{T_jT_k} - \rho_{T_iT_j}\rho_{T_iT_k} ,$$

which, given equation (3), equals:

$$\rho_{T_iT_j}\rho_{T_iT_k}\left(\frac{\rho_{ii}}{a_i^2} - 1\right) .$$

Considering cases where a single factor reproduces the intercorrelations among X_1 , X_2 , and X_3 and $0 < a_i^2 < 1$ $(i = 1,2,3)$, the above expression can be seen to have the following implications:

A. When $\rho_{T_iT_j}$ and $\rho_{T_iT_k}$ have the same sign,

1. $a_i^2 < \rho_{ii}$ implies $\rho_{T_jT_k \cdot T_i} > 0$,

2. $a_i^2 = \rho_{ii}$ implies $\rho_{T_jT_k \cdot T_i} = 0$,

3. $a_i^2 > \rho_{ii}$ implies $\rho_{T_jT_k \cdot T_i} < 0$;

B. When $\rho_{T_iT_j}$ and $\rho_{T_iT_k}$ have opposite signs,

1. $a_i^2 < \rho_{ii}$ implies $\rho_{T_jT_k \cdot T_i} < 0$,

2. $a_i^2 = \rho_{ii}$ implies $\rho_{T_jT_k \cdot T_i} = 0$,

3. $a_i^2 > \rho_{ii}$ implies $\rho_{T_jT_k \cdot T_i} > 0$.

These results show that when the correlations among the observed scores are reproduced by a single factor with squared loadings between 0 and 1, no conclusions are warranted regarding the partial correlations among the true scores. Given positive reliabilities and nonzero intercorrelations among observed scores, if the three observed variables do not fit the single factor model, then the three partial correlations among true scores may be positive or negative but not zero.

The relationship between the observed loadings on a single common factor, the partial correlations among observed scores, and the partial correlations among true scores may be clarified by the example depicted in Figure 2. For the case $\rho_{T_1T_2} = .50$ and $\rho_{11} = \rho_{22} = \rho_{33} = .50$, Figure 2 shows the possible values of $\rho_{T_1T_3}$ and $\rho_{T_2T_3}$. A set of regions is defined within which the factor loadings on a single common factor, the partial correlations among observed scores, and the partial correlations among true scores have specified characteristics. The ellipse in Figure 2 contains the values for which the determinant of the matrix containing the intercorrelations of T_1, T_2, and T_3 is greater than or equal to zero. Larger values of $\rho_{T_1T_2}$ would define a thinner ellipse and smaller values a rounder ellipse. The numbers inside the ellipse identify the various regions of the ellipse, and the letters identify line segments separating regions. For the regions in Figure 2, the factor loadings (a_i) for a single common factor that will reproduce the intercorrelations among the observed scores, the partial correlations among the observed scores $(\rho_{jk.i})$, and the partial correlations among the true scores $(\rho_{T_jT_k.T_i})$ are shown in Table 1. The values of a_i, $\rho_{jk.i}$

Insert Figure 2 about here

----

and $\rho_{T_jT_k \cdot T_i}$ for values of $\rho_{T_iT_j}$ and $\rho_{T_iT_k}$ on the boundaries between regions are shown in Table 2.

Insert Tables 1 and 2 about here

As was stated in implications A.2 and B.2 above, $\rho_{T_jT_k \cdot T_i}$ equals zero when $a_i^2 = \rho_{ii}$. This occurs on line segments co, do, io, jo, cmd, and inj. When $a_i^2 = 1$ (line segments bo, eo, ho, and ko) the partial, $\rho_{jk.i}$, among observed scores is zero; however, $\rho_{T_jT_k \cdot T_i}$ is nonzero. The location of line boh and line eok depends on the magnitude of ρ_{11} and ρ_{22} : boh is defined by points where $\rho_{T_2T_3} = \rho_{11}\rho_{T_1T_2}\rho_{T_1T_3}$ and eok is defined by points where $\rho_{T_1T_3} = \rho_{22}\rho_{T_1T_2}\rho_{T_2T_3}$. A line where $a_3^2 = 1$ does not exist for this example because there are no possible values of $\rho_{T_1T_3}$ and $\rho_{T_2T_3}$ for which $\rho_{T_1T_2}$ equals $\rho_{33}\rho_{T_1T_3}\rho_{T_2T_3}$. Regions 2a, 3a, 4, 6a, 7a, and 8 are of interest since they define combinations of $\rho_{T_iT_j}$ and $\rho_{T_iT_k}$ for which a partial correlation for observed scores and a partial correlation for true scores have opposite signs. Regions 1, 2a, 3a, 4, 5, 6a, 7a, and 8 are where a satisfactory single-factor solution is obtained yet all three correlations between pairs of true scores with the third true score partialed out are nonzero. Different conclusions about the number of underlying conceptual variables involved in the phenomenon presumably would be drawn for instances in those regions.

This problem should not be dealt with by simply invoking the principle of parsimony and thereby concluding that the fit of a single factor model indicates that there is only one dimension underlying the phenomenon. Rather, the problem should be dealt with by obtaining the additional information that is necessary to make inferences within a given model. A brief discussion of the use of multiple measures to obtain the needed information is presented below in the section on needed additional information.

Errors of Measurement in the Analysis of Covariance

Campbell and Erlebacher (1970) have provided a much needed criticism of the common misuse of the analysis of covariance as a means of trying to adjust for preexisting differences between experimental and control groups for the evaluation of compensatory education programs. They argue that "error" and "uniqueness" in the covariate result in bias when the groups differ on the direction of underestimating the slope of the regression of the dependent variable, on the covariate (for a good discussion see Cochran, 1968). Porter (1967) has illustrated the nature of the resulting bias for various group differences in means on the covariate and on the dependent variable. When using the analysis of covariance, bias due to errors of measurement in the covariate might make a compensatory education program look bad (or good).

The effect of "uniqueness" depends on its sources. If uniqueness is due to errors of validity (e.g., a perfectly reliable symptom of the underlying variable), then bias will result in the same way that it does from unreliability. On the other hand, if uniqueness merely refers to unshared variance between the covariate and the dependent variable as in Campbell and Erlebacher's (1970) treatment of covariance adjustments, then the question of bias is ambiguous. Given independent errors, unshared variance may be due to unreliability, invalidity or a lack of perfect correlation between underlying variables. The latter is not a source of bias and should not be corrected for as is done by Campbell and Erlebacher's adjustment procedure.

This problem needs to be viewed from the perspective of Lord's (1967) paradox. Lord has shown that the comparison of preexisting groups by means of an analysis of covariance (statistician 2) and by means of an analysis of difference scores (statistician 1) can result in paradoxically different results, both of which are manifestly correct. In his hypothetical illustrative, example, Lord depicted an experiment in which girls received one diet and boys another. For each group the mean and variance of the final weight was identical to the mean and variance of the initial weight. There were preexisting differences between the groups in mean weight, and for each group the within-group correlation between initial and final weight was .50. Assuming that the weight measures are error free, the above correlation would be the correlation between true initial weight and true final weight. In the absence of measurement errors the analysis of mean change would indicate no "treatment" effect, whereas the analysis of covariance would indicate a "treatment" effect.

Campbell and Erlebacher (1970) have suggested that in pretest-posttest designs a "common-factor coefficient" might be used to correct for errors of measurement and uniqueness in the covariate. Using the proper common factor coefficients for both pretest and posttest in the standard correction for attenuation formula would result in a "corrected" pretest-posttest correlation of 1.00. Assuming equal coefficients for the pretest and the posttest, the common factor coefficient for Lord's example would be .50. Applying this "correction" would increase the slope of the within-group regression lines to 1.00 and result in identical intercepts for the two groups. In essence, Campbell and Erlebacher have devised a roundabout way of siding with Lord's first statistician. However, they have not resolved Lord's paradox. Rather than impose a restriction, such as the one that the "corrected" correlation between pretest and posttest be 1.00 (which, in our opinion, is unjustified), it would seem far better to conclude with Lord (1967) that ". . . there simply is no logical or statistical procedure that can be counted on to make proper allowances for uncontrolled preexisting differences between groups" (p. 305).

Needed Additional Information for Fallible Measures

Dealing with fallible measures will generally require additional assumptions and additional information. In some instances, using parallel forms of one or more of the measures may provide the needed additional information. One difficulty with this procedure is that most observed measures are really symptoms or indirect measures of the variable or influence to be measured, which is to say that even if the symptoms were measured with perfect reliability, they would be imperfectly correlated with the "true" variable. The researcher must decide which symptoms are reflections of the relevant underlying variable. This question is crucial since different sets of symptoms will typically define different "true" factors depending on the particular statistical procedure employed. The multitrait-multimethod approach introduced by Campbell and Fiske (1959) attempts to deal with this validity problem by using different methods of measuring the same variable. Correlations between different method measures of the same trait typically will correlate less than equivalent measures, i.e., in this model the classical psychometric approach using parallel forms is apt to underestimate correlations among underlying conceptual variables. An alternative way of stating this problem is to assume that part of the correlation between the two measures X_1 and X_1^* of T_1 is due to correlated errors of measurement and that factors causing this correlation are uncorrelated with the true scores. In this case, the square root of the correlation between X_1 and X_1^* no longer provides a reasonable estimate of the correlation between X_1 and T_1. Assuming that the errors are positively correlated, the correlation between X_1 and X_1^* will overestimate the squared correlation between X_1 and T_1 and using this inflated coefficient to correct for attenuation will result in the kind of undercorrection that Brewer et al. (1970) warned against. Correlated errors may, in fact, be one of the reasons that Brewer et al. wanted to correct for "uniqueness." There are advantages, however, to formulating the problem in terms of correlated errors rather than simply saying that we should correct for uniqueness. The former makes it possible to devise procedures for estimating the desired coefficient (the correlation between X_1 and T_1) given the possibility of either positively or negatively correlated errors, whereas the latter only allows the conclusion that the correlation between X_1 and X_1^* overestimates the desired coefficient *if the errors are in fact positively correlated*.

Conclusion

From our perspective, "focusing on the conceptual problem of choosing a one-factor vs. a two-factor model" (Brewer et al., 1970, p. 3) distracts the researcher's attention from the task of constructing a model which is consistent with everything we know or hypothesize about the phenomena under study. Any inferences will necessarily be no more valid than the assumptions made about reality. For heuristic purposes we have assumed that the linear additive model was relevant; however, there is no rule of nature that effects are either linear or additive. No provision was made, e.g., for catalytic, feedback, or interactional type influences. It is important for the research design to be set up to study the question of which of the plausible alternative models more closely simulates reality. Rather than focus on the conceptual problem of choosing a one-factor vs. a two-factor model, it seems to us far more worthwhile to spend time in designing the study to explore the relevant alternate models, ensuring collection of the information necessary to test which is the best simulation of reality. Depending on the problem, the factor model may be one of the alternatives. The assumption that the factor model is a priori relevant appears to us to be unjustified given the current state of the art.

References

Brewer, M. B., Campbell, D. T., & Crano, W. D. Testing a single-factor model as an alternative to the misuse of partial correlations in hypothesis-testing research. Sociometry, 1970, 33, 1-11.

Campbell, D. T., & Erlebacher, A. How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful. In J. Hellmuth (Ed.), Compensatory education--a national debate. Vol. 3, Disadvantaged Child. New York: Brunner Mazel, Inc. 1970.

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

Cochran, W. G. Errors of measurement in statistics. Technometrics, 1968, 10, 637-666.

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.

Lord, F. M. A paradox in the interpretation of group comparisons. Psychological Bulletin, 1967, 68, 305-306.

Porter, A. C. The effects of using fallible variables in the analysis of covariance. (Doctoral dissertation, University of Wisconsin) Ann Arbor, Mich.: University Microfilms, 1967. No. 67-12, 147.

Footnotes

[1] The research reported herein was performed pursuant to Grant No. OEG-2-700033 (509) with the United States Department of Health, Education, and Welfare and the Office of Education.

[2] We are grateful to Frederic M. Lord for suggesting the idea that was used for the illustrative example in Figure 1.

Table 1

Values of Factor Loadings and Partial Correlations

for Regions of Figure 2

	Factor Loadings			Partial Correlations Among Observed Scores			Partial Correlations Among True Scores		
Region	a_1	a_2	a_3	$\rho_{23.1}$	$\rho_{13.2}$	$\rho_{12.3}$	$\rho_{T_2T_3.T_1}$	$\rho_{T_1T_3.T_2}$	$\rho_{T_1T_2.T_3}$
1	+	+	+	+	+	+	+	+	+
2a	+	+	+	+	+	+	-	+	+
2b	G[a]	+	+	-	+	+	-	+	+
3a	+	+	+	+	+	+	+	-	+
3b	+	G	+	+	-	+	+	-	+
4	+	+	+	+	+	+	+	+	-
5	-	-	+	-	-	+	-	-	+
6a	-	-	+	-	-	+	+	-	+
6b	G	-	+	+	-	+	+	-	+
7a	-	-	+	-	-	+	-	+	+
7b	-	G	+	-	+	+	-	+	+
8	-	-	+	-	-	+	-	-	-
9	i	i	i	-	+	+	-	+	+
10	i	i	i	+	-	+	+	-	+

[a] G denotes that the factor loading is greater than 1.0 in absolute value.

Table 2

Values of Factor Loadings and Partial Correlations

for Lines Separating Regions in Figure 1

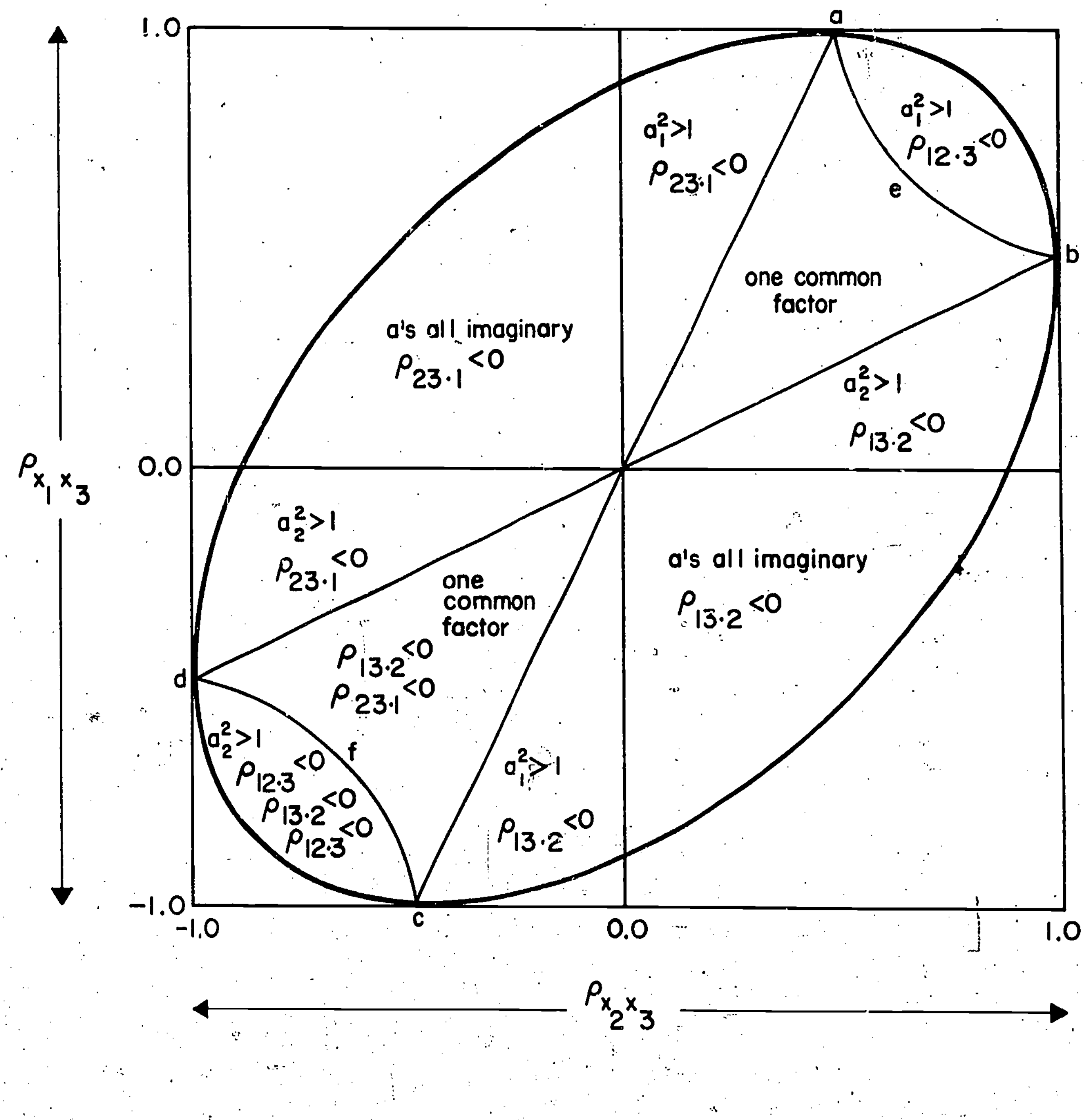
Line Segment	Factor Loadings			Partial Correlations Among Observed Scores			Partial Correlations Among True Scores		
	a_1	a_2	a_3	$\rho_{23.1}$	$\rho_{13.2}$	$\rho_{12.3}$	$\rho_{T_2T_3 \cdot T_1}$	$\rho_{T_1T_3 \cdot T_2}$	$\rho_{T_1T_2 \cdot T_3}$
ao	U[a]	0	0	-	+	+	-	+	+
bo	1	+	+	0	+	+	-	+	+
co	$\sqrt{\rho_{11}}$	+	+	+	+	+	0	+	+
do	+	$\sqrt{\rho_{22}}$	+	+	+	+	+	0	+
eo	+	1	+	+	0	+	+	-	+
fo	0	U	0	+	-	+	+	-	+
go	U	0	0	+	-	+	+	-	+
ho	-1	-	+	0	-	+	+	-	+
io	$-\sqrt{\rho_{11}}$	-	+	-	-	+	0	-	+
jo	-	$-\sqrt{\rho_{22}}$	+	-	-	+	-	0	+
ko	-	-1	+	-	0	+	-	+	+
eo	0	U	0	-	+	+	-	+	+
cmd	+	+	$\sqrt{\rho_{33}}$	+	+	+	+	+	0
inj	-	-	$\sqrt{\rho_{33}}$	-	-	+	-	-	0

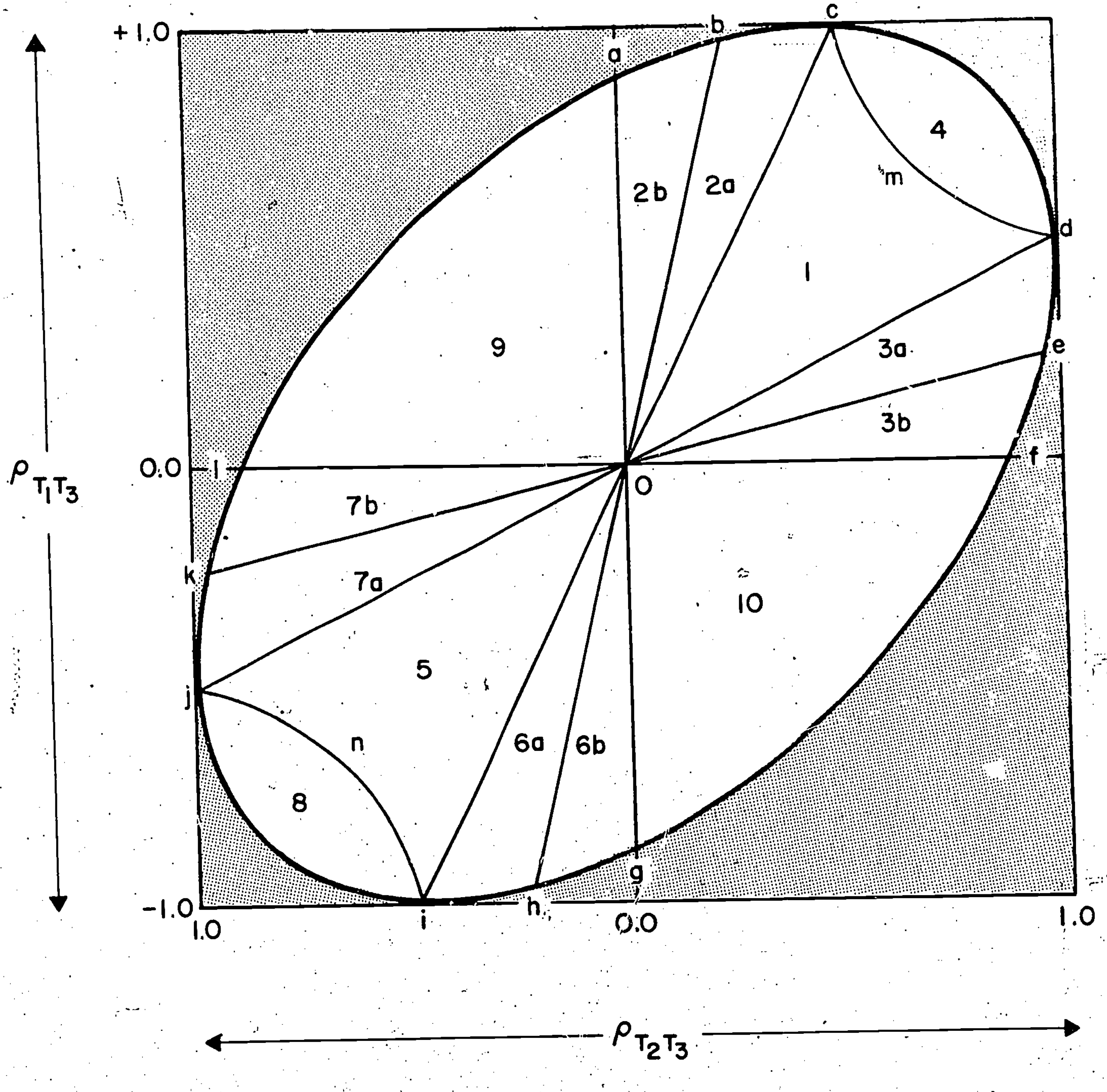
[a] U denotes that the factor loading is undefined.

Figure Captions

Fig. 1. Regions which define values of factor loading and partial correlations for possible values of $\rho_{X_1X_3}$ and $\rho_{X_2X_3}$ given $\rho_{X_1X_2} = .50$.

Fig. 2. Regions which define values of factor loadings and partial correlations for possible values of $\rho_{T_1T_3}$ and $\rho_{T_2T_3}$ given $\rho_{T_1T_2} = .50$, and $\rho_{11} = \rho_{22} = \rho_{33} = .50$.


1.0
a
$a_1^2>1$
$\rho_{23\cdot1}<0$
$a_1^2>1$
$\rho_{12\cdot3}<0$
e
b
one common factor
a's all imaginary
$\rho_{23\cdot1}<0$
$a_2^2>1$
$\rho_{13\cdot2}<0$
$\rho_{x_1x_3}$
0.0
$a_2^2>1$
$\rho_{23\cdot1}<0$
one common factor
a's all imaginary
$\rho_{13\cdot2}<0$
$\rho_{13\cdot2}<0$
$\rho_{23\cdot1}<0$
d
$a_2^2>1$
$\rho_{12\cdot3}<0$
f
$\rho_{13\cdot2}<0$
$\rho_{12\cdot3}<0$
$a_1^2>1$
$\rho_{13\cdot2}<0$
-1.0
-1.0
c
0.0
1.0
$\rho_{x_2x_3}$


+1.0
0.0
-1.0
$\rho_{T_1T_3}$
1.0
0.0
1.0
$\rho_{T_2T_3}$
a
b
c
d
e
f
g
h
i
j
k
l
m
n
O
1
2a
2b
3a
3b
4
5
6a
6b
7a
7b
8
9
10

Identification and Estimation in Path Analysis with Unmeasured Variables

Abstract

A variety of path models involving unmeasured variables are formulated in terms of Jöreskog's (1970a) general model for the analysis of covariance structures.

Identification and Estimation in Path Analysis with Unmeasured Variables*

A variety of authors (e.g., Blalock, 1969; Costner, 1969; Heise, 1969) have applied path analysis to problems involving multiple indicators of underlying constructs. An important and often algebraically complex feature of such analysis is the determination of identifiability of model parameters. The purpose of this discussion is to demonstrate how a visual inspection of the path diagram can be used to simplify the identification question and how these problems may be formulated in Jöreskog's (1970a) general model.

I. A Single Factor Model

Consider the case of a single underlying factor (F_1) with three observed measures $(X_1, X_2,$ and $X_3)$ as shown in Figure 1.a. The factor loadings $(\rho_{X_iF_1})$ in this model equal the standardized path coefficients $(b_1^*, b_2^*,$ and $b_3^*)$, given the assumption that the residuals $e_1, e_2,$ and e_3 are independent of each other and of the factor. It is convenient, though not necessary, to assume that both measured and unmeasured variables are standardized. For heuristic purposes observed correlations will be designated with "r" and expected values of these correlations by "ρ". The expected correlations will differ from the corresponding observed correlations because of sampling and model specification errors.

*The research reported herein was performed pursuant to Grant No. OEG-2-700033(509) with the United States Department of Health, Education, and Welfare, and the Office of Education.

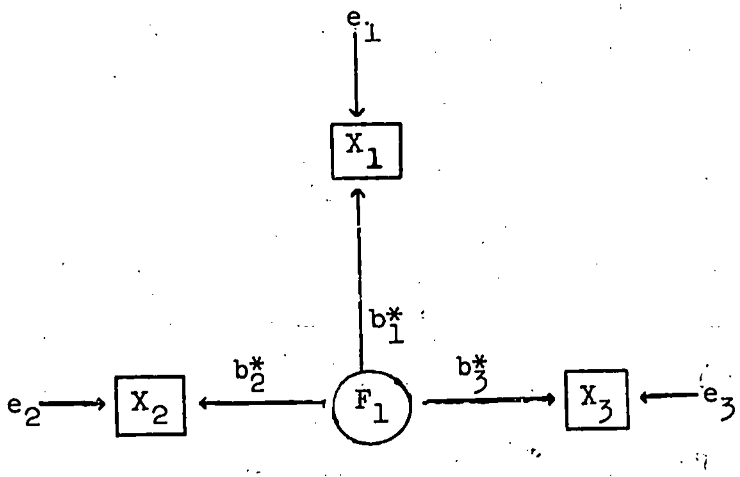


Fig. 1.a. A Single Factor Model

A path analysis of this model yields the equations:

$$\rho_{12} = b_1^* b_2^* \ ,$$

$$\rho_{13} = b_1^* b_3^* \ , \qquad (1)$$

and $$\rho_{23} = b_2^* b_3^* \ .$$

Assuming nonzero correlations, equations (1) yield:

$$(b_1^*)^2 = \frac{\rho_{12}\rho_{13}}{\rho_{23}} = \rho^2_{X_1F_1} \ ,$$

$$(b_2^*)^2 = \frac{\rho_{12}\rho_{23}}{\rho_{13}} = \rho^2_{X_2F_1} \quad \text{and} \qquad (2)$$

$$(b_3^*)^2 = \frac{\rho_{13}\rho_{23}}{\rho_{12}} = \rho^2_{X_3F_1} \ .$$

Given only three observed measures the model is just identified, i.e., the observed and expected correlations are identical. With more than three measures

$$\rho^2_{X_iF_1} = (b_i^*)^2 = \frac{\rho_{ij}\rho_{ik}}{\rho_{jk}} \ , \quad \text{where} \quad i \neq j \neq k \quad \text{and} \qquad (2a)$$

assuming $\rho_{jk} \neq 0$. If there were a causal linkage (e.g., $F_1 \rightarrow I_1 \rightarrow I_2 \rightarrow X_i$) from F_1 to X_i then $\rho_{X_iF_1}$ would be the product of the intervening path coefficients, i.e., the product of the path coefficients in the chain from F_1 to X_i would be identified. If any loading exceeded unity, the model would be rejected. When there are $m > 3$ observed measures then the loadings will be overidentified. The number of overidentifying restrictions is simply the number of distinct correlations $m(m - 1) \div 2$ less the number (m) of $\rho_{X_iF_1}$ to be estimated. Maximum likelihood or least squares estimates for overidentified models can be obtained using Jöreskog's (1970a) general method for the analysis of covariance structures. We use path analysis only to study the identifiability problem, not for estimation purposes (Hauser & Goldberger, 1970; Werts, Jöreskog, & Linn, in press).

The above analysis leads to our "rule of three": Whenever the correlations among at least three observed variables may be completely ascribed to the presence of an underlying factor, then the loadings (correlations) for each observed variable on that factor are identifiable. An important qualification is that the expected correlation between any two observed variables cannot be zero since equation (2a) would not be defined when that correlation was in the denominator. In practice, small expected correlations may lead to unstable parameter estimates, i.e., highly unreliable measures result in unreliable parameter estimates.

II. Generalizations

The Figure 1.a. model with or without intervening, unmeasured variables going from F_1 to X_i is too limited for most causal analyses. Our purpose in this section is to consider other causal patterns which satisfy the "rule of three," i.e., in which the observed correlations among three variables are nonzero and may be ascribed to the presence of an underlying factor. Equations (1), and therefore (2), would still hold if for one of the measures (e.g., X_1) $X_1 \to F_1$ and the residual θ_1 of this regression of F_1 on X_i were independent of the other residuals e_2 and e_3 , as shown in Figure 1.b.

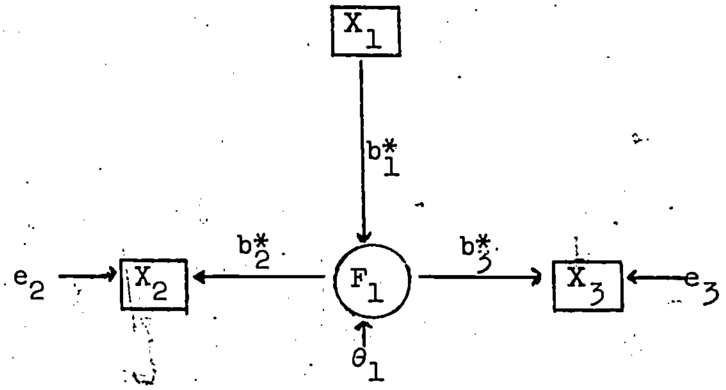


Figure 1.b.

If two observed measures influence F_1 , e.g., $X_1 \to F_1$ and $X_2 \to F_1$ then it is no longer true that the correlation between these measures equals the product of the corresponding path coefficients, e.g., ρ_{12} would not in general equal $b_1^* b_2^*$.

Given that all residuals are independent, when there is an intervening variable (I_1) between X_i and F_1, the correlation between a pair of observed variables X_i and X_j will equal the product of the intervening path coefficients when $X_i \leftarrow I_1 \leftarrow F_1 \to X_j$, $X_i \leftarrow I_1 \to F_1 \to X_j$, $X_i \to I_1 \to F_1 \to X_j$, and $X_i \leftarrow I_1 \leftarrow F_1 \leftarrow X_j$; but not when two arrows point towards the same variable, e.g., $X_i \to I_1 \leftarrow F_1 \to X_j$ or $X_i \to I_1 \to F_1 \leftarrow X_j$. In general the correlation between two observed variables may be stated as the product of the intervening path coefficients whenever the causal linkage between these variables does not include a variable which is caused by two other variables, i.e., when two causal arrows point towards a variable. To identify the loadings on a factor we need to find three observed variables which are causally linked through that factor, the linkages satisfying the above criteria.

III. Examples

A. Our first example, which corresponds to Figure 1 in Wiley and Wiley (1970), is shown in Figure 2.a.

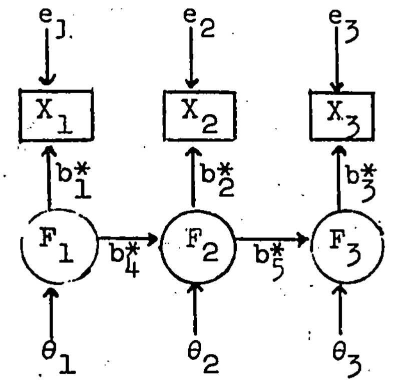


Figure 2. a.

Tracing linkages for F_2 :

$X_1 \leftarrow F_1 \rightarrow F_2 \rightarrow F_3 \rightarrow X_3$,

$X_1 \leftarrow F_1 \rightarrow F_2 \rightarrow X_2$, and

$X_2 \leftarrow F_2 \rightarrow F_3 \rightarrow X_3$.

Since these three linkages all include F_2 and satisfy the requirements of the "rule of three" we may conclude that the factor loadings $(\rho_{X_iF_2})$, i.e., the correlations of each observed variable with F_2 , are identified. Thus,

$$\rho_{X_1F_2} = b_1^*b_4^* ,$$

$$\rho_{X_2F_2} = b_2^* , \text{ and } \qquad (3)$$

$$\rho_{X_3F_2} = b_3^*b_5^* .$$

The factor loadings on F_1 are not identified because the correlation between X_2 and X_3 cannot be completely ascribed to F_1. Likewise the loadings on F_3 are not identified because the correlation between X_1 and X_2 cannot be ascribed to F_3. Jöreskog (1970b) shows that this model may be estimated by a single factor model with F_2 as the common factor and that the example may be generalized to more than three measured variables.

B. Our second example (see Figure 2.b) corresponds to Figure 4 in Costner (1969). The analysis is identical whether $F_1 \rightarrow F_2$ or $F_2 \leftarrow F_1$.

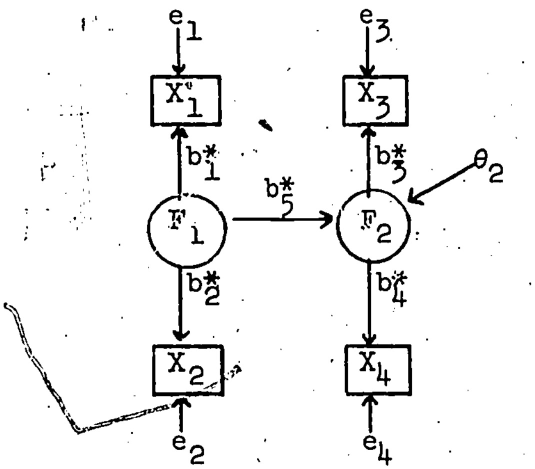


Figure 2.b.

Tracing linkages:

$X_1 \leftarrow F_1 \rightarrow X_2$, (4a)

$X_1 \leftarrow F_1 \rightarrow F_2 \rightarrow X_3$, (4b)

$X_1 \leftarrow F_1 \rightarrow F_2 \rightarrow X_4$, (4c)

$X_2 \leftarrow F_1 \rightarrow F_2 \rightarrow X_3$, (4d)

$X_2 \leftarrow F_1 \rightarrow F_2 \rightarrow X_4$, and (4e)

$X_3 \leftarrow F_2 \rightarrow X_4$. (4f)

For F_1 the factor loadings may be identified by linkages 4a,b,d or by 4a,c,e, i.e., these loadings are overidentified and

$$\rho_{X_1F_1} = b^*_1 ,$$

$$\rho_{X_2F_1} = b^*_2 ,$$

$$\rho_{X_3F_1} = b^*_3 b^*_5 , \text{ and}$$

$$\rho_{X_4F_1} = b^*_4 b^*_5 .$$

The factor loadings for F_2 may be identified by 4b,c,f or 4d,e,f and:

$$\rho_{X_1F_2} = b^*_1 b^*_5 ,$$

$$\rho_{X_2F_2} = b^*_2 b^*_5 ,$$

$$\rho_{X_3F_2} = b^*_3 , \text{ and}$$

$$\rho_{X_4F_2} = b^*_4 .$$

Since b^*_1 and b^*_2 are identified, b^*_5 is also identified by these equations.

The analysis may be complicated by assuming e_1 correlated with e_3, in which case linkage 4b would not be valid, however the conditions of the "rule of three" would still be satisfied for F_1 and F_2 and all path coefficients and correlations between errors are (just) identified. Such a model would correspond to Figure 5.a. in Costner (1969).

C. The next example, corresponding to Figure 1 in Blalock (1963), is shown in Figure 2.c. This model is basically a variation on the model of Figure 1.b.

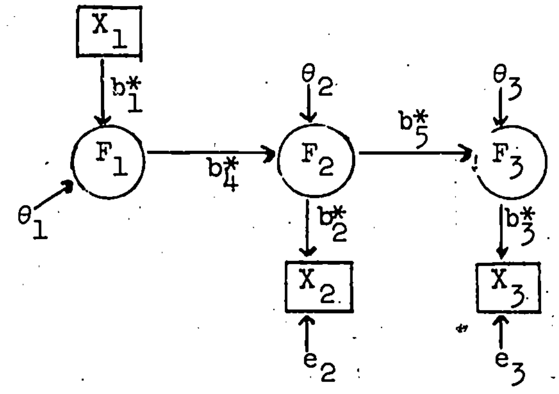


Figure 2.c.

This model differs from that in Figure 2.a. in that $X_1 \rightarrow F_1$ instead of $F_1 \rightarrow X_1$. The linkages are:

$X_1 \rightarrow F_1 \rightarrow F_2 \rightarrow X_2$,

$X_1 \rightarrow F_1 \rightarrow F_2 \rightarrow F_3 \rightarrow X_3$, and

$X_2 \leftarrow F_2 \rightarrow F_3 \rightarrow X_3$.

Since F_2 is in all three linkages which satisfy the "rule of three," the factor loadings for F_2 are identified and

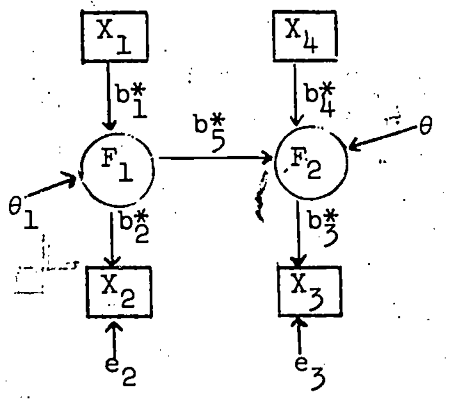
$$\rho_{X_1F_2} = b_1^* b_4^* = \sqrt{r_{12}r_{13} \div r_{23}} \quad , \tag{5a}$$

$$\rho_{X_2F_2} = b_2^* = \sqrt{r_{12}r_{23} \div r_{13}} \quad , \text{ and} \tag{5b}$$

$$\rho_{X_3F_2} = b_3^* b_5^* = \sqrt{r_{13}r_{23} \div r_{12}} \quad . \tag{5c}$$

Since r_{12} cannot be ascribed to F_3 and r_{23} cannot be ascribed to F_1 , the loadings on these factors are not identified. Our heuristic device would have been helpful to Blalock (1963) since he obtained the equations corresponding to the linkages shown above, but did not solve them for the equivalent of equations 5a, b, and c.

D. Our fourth example, shown in Figure 2.d., corresponds to Figure 2 in Blalock (1963).



θ = residual of F_2 on X_4 and F_1 .

Figure 2.d.

Tracing linkages which satisfy our rule:

$$X_1 \rightarrow F_1 \rightarrow X_2 \quad , \tag{6a}$$

$$X_1 \rightarrow F_1 \rightarrow F_2 \rightarrow X_3 \quad , \tag{6b}$$

$$X_2 \leftarrow F_1 \rightarrow F_2 \rightarrow X_3 \quad , \text{ and} \tag{6c}$$

$$X_4 \rightarrow F_2 \rightarrow X_3 \quad . \tag{6d}$$

In this model it is assumed that X_4 is independent of X_1 and X_2 . The loadings on F_1 are identified by linkages 6a,b and c and therefore:

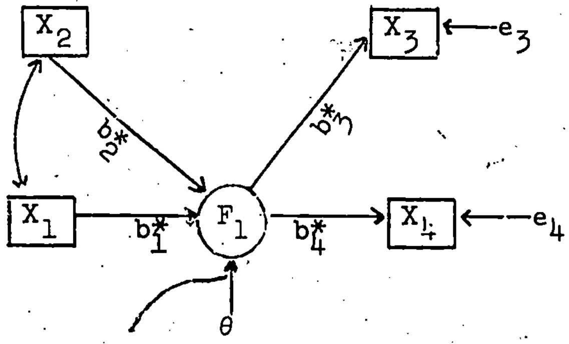
$$\rho_{X_1F_1} = b^*_1 \quad , \tag{7a}$$

$$\rho_{X_2F_1} = b^*_2 \quad , \text{ and} \tag{7b}$$

$$\rho_{X_3F_1} = b^*_3 b^*_5 \quad . \tag{7c}$$

It is not possible to find three observed variables whose linkages satisfy our rule for F_2 , i.e., the linkage between X_1 and X_4 has two arrows pointing at F_2 and the linkage between X_1 and X_2 does not include F_2. Since $\rho_{34} = b_3^* b_4^*$ it follows from equation (7c) that $\rho_{X_3F_1} b_4^* = \rho_{34} b_5^*$.

E. The fifth example, shown in Figure 2.e., has the special feature of two observed nonindependent variables influencing an unobserved variable. It corresponds to Figure 4 in Blalock (1969).



θ = residual of F_1 on X_1 and X_2 regression.

Figure 2.e.

When X_2 is deleted X_1 , X_3, and X_4 form the model in Figure 1.b. from which we conclude that the correlations of X_1 , X_3 , and X_4 with F_1 are identified. Similarly when X_1 is deleted the correlations of X_2 , X_3, and X_4 with F_1 are identified. Given the correlations among X_1 , X_2, and F_1 the path coefficients b_1^* and b_2^* may be identified since:

$$b_1^* = \frac{\rho_{X_1F_1} - \rho_{12}\rho_{X_2F_1}}{1 - \rho_{12}^2} \quad \text{and}$$

$$b_2^* = \frac{\rho_{X_2F_1} - \rho_{12}\rho_{X_1F_1}}{1 - \rho_{12}^2} \quad .$$

F. Our last example, shown in Figure 2.f. corresponds to Figure 9.b. in Costner (1969)

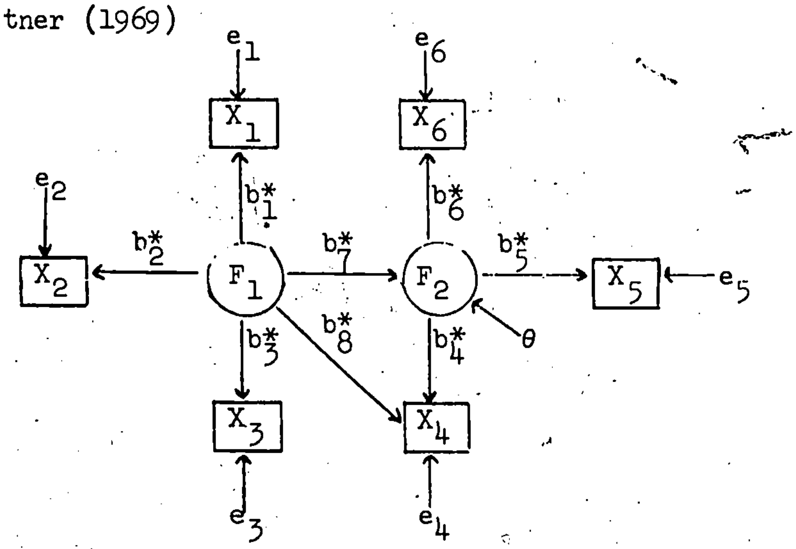


Figure 2.f.

From the analysis of the Figure 2.d. model we may deduce that when X_4 is excluded that b_1^* , b_2^* , b_3^* , b_5^* , b_6^* , and b_7^* are identified. Using the variables X_1 , X_2 , and X_4 we know from our analysis of the Figure 1 model that the correlation of X_4 with F_1 ($\rho_{X_4 F_1}$) is identified and similarly using X_4 , X_5 , and X_6 we know that the correlation of X_4 with F_2 ($\rho_{X_4 F_2}$) is identified. Since the correlations among F_1 , F_2 , and X_4 are identified it follows that the path coefficients b_4^* and b_8^* , which are functions of these correlations, are identified. As compared to Costner's (1969) rather complex algebraic analysis of this problem, it may be seen that we are satisfied in merely knowing that the model parameters are identified.

IV. Estimation

Jöreskog's (1970a) general model for the analysis of covariance structures can be used to estimate the parameters for the models discussed above. Werts, Jöreskog and Linn (in press) discuss the use of Jöreskog's model from the perspective of path analysis. Use of the associated computer program (Jöreskog, Gruvaeus, & van Thillo, 1970) for the present purposes requires the investigator to specify a matrix Λ corresponding to the factor loadings in factor analysis; a matrix Φ which is the variance-covariance matrix of the unmeasured factors, and a matrix Θ of residual variances. The matrices B and Ψ in Jöreskog's formula are taken as the identity and zero matrix respectively.

Consider for example the model in Figure 1 in which

$$\Lambda = \begin{bmatrix} b_1^* \\ b_2^* \\ b_3^* \end{bmatrix} ,$$

$$\Phi = [1] ,$$

and

$$\Theta = \begin{bmatrix} V_{e_1} & 0 & 0 \\ 0 & V_{e_2} & 0 \\ 0 & 0 & V_{e_3} \end{bmatrix} .$$

Define: $\underset{\sim}{X}$ = column vector of standardized observed variables,

$\underset{\sim}{F}$ = column vector of factors, and

$\underset{\sim}{e}$ = column vector of residuals.

In matrix terminology:

$$\underset{\sim}{X} = \Lambda\underset{\sim}{F} + \underset{\sim}{e} \quad . \qquad (8)$$

Equation (8) is shorthand for the path equations (all variables standardized):

$X_1 = b_1^*F_1 + e_1$,

$X_2 = b_2^*F_1 + e_2$, and

$X_3 = b_3^*F_1 + e_3$.

It can be seen that Λ is the matrix of the coefficients of F_1 . The parameters in the matrices specifying the model structure in Jöreskog's model are of three kinds: (1) fixed parameters that have been assigned given values; (2) constrained parameters that are unknown but equal to one or more other parameters; and (3) free parameters that are unknown and not constrained to be equal to any other parameter. In the above example the unity in Φ is a fixed parameter, whereas the b_i^* in Λ and the V_{e_i} in Θ are free parameters.

The expected variance-covariance matrix Σ for this problem is:

$$\Sigma = \Lambda\Phi\Lambda' + \Theta^2 \qquad (9)$$

where the 1 in Φ is the variance of F_1 , for convenience standardized (i.e., equal to unity) and Θ^2 is a diagonal matrix whose elements are the error variances (V_{e_i}) . Equation (9) should be recognized as a shorthand way of expressing all the path equations relating expected model correlations to model parameters, i.e.,

$$\Sigma = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{bmatrix} ,$$ (where unities indicate observed variables were standardized).

Equation (9) states:

$$1 = (b_1^*)^2 + V_{e_1} \quad ,$$

$$1 = (b_2^*)^2 + V_{e_2} \quad ,$$

$$1 = (b_3^*)^2 + V_{e_3} \quad ,$$

$$\rho_{12} = b_1^* b_2^* \quad ,$$

$$\rho_{13} = b_1^* b_3^* \quad , \text{ and}$$

$$\rho_{23} = b_2^* b_3^* \quad .$$

This short description for a single model contrasts with the path analysis approach to estimation used by Costner (1969) and Blalock (1969) in the following respects:

(a) The matrix Σ of expected correlations between observed variables will differ from the actually observed matrix of correlations because of sampling and/or model specification errors. Thus we do not use observed correlations in our equations as in the usual path analysis approach. Instead, Jöreskog's program attempts to minimize the difference between observed and expected variance-covariance matrices using either a least squares or maximum likelihood approach. In large samples, assuming that observed variables are distributed normally, a chi square statistic is produced which measures the overall fit of the model to the data. Another way of gauging fit is to compare the differences between the observed and expected correlations generated by the model.

(b) The degrees of freedom (df) for the χ^2 measure are equal to the number of overidentifying restrictions. In path analysis this corresponds to the number of different ways the path equations may be solved for each parameter. To compute the df it is necessary to count the number of distinct elements in Σ (i.e., $m(m+1) \div 2$) and subtract the number of parameters to be estimated (e.g., $b_1^*, b_2^*, b_3^*, V_{e_1}, V_{e_2}$, and V_{e_3}). There is no need to solve the path equations in Jöreskog's approach, although the identifiability must be known.

To analyze the model in Figure 1.b., we merely need to note that when X_1 and F_1 are standardized the regression of X_1 on F_1 equals that of F_1 on X_1 and the residuals are identical. Thus we may use the same estimation procedure for this model as for that in Figure 1.a. (where $\theta_1 = e_1$). Likewise the models in Figures 2.a. and 2.c. may be estimated by ignoring F_1 and F_3 and treating X_1, X_2, and X_3 as indicators of the common factor F_2.

The model in Figure 2.b. with the added feature of e_1 and e_3 correlated requires special treatment. The equations are:

$$X_1 = b_1^* F_1 + e_1 \quad ,$$
$$X_2 = b_2^* F_1 + e_2 \quad ,$$
$$X_3 = b_3^* F_2 + e_3 \quad ,$$
$$X_4 = b_4^* F_2 + e_4 \quad , \text{ and}$$
$$F_2 = b_5^* F_1 + \theta_2 \quad .$$

We know that b_5^* is equal to the correlation between F_1 and F_2 so there is no need to replace F_2 by F_1 and θ_2 in the first four equations. To specify a correlation between e_1 and e_3, all residuals must be treated as factors, i.e., $\underset{\sim}{F}' = (F_1, F_2, e_1, e_2, e_3, e_4)$. The structure is:

$$\Lambda = \begin{bmatrix} b_1^* & 0 & b_{e_1}^* & 0 & 0 & 0 \\ b_2^* & 0 & 0 & b_{e_2}^* & 0 & 0 \\ 0 & b_3^* & 0 & 0 & b_{e_3}^* & 0 \\ 0 & b_4^* & 0 & 0 & 0 & b_{e_4}^* \end{bmatrix} ,$$

and

$$\Phi = \begin{bmatrix} 1 & b_5^* & 0 & 0 & 0 & 0 \\ b_5^* & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & \rho_{e_1e_3} & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & \rho_{e_1e_3} & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} .$$

In contrast to previous formulations the error variances are standardized so that the correlations between e_1 and e_3 and F_1 and F_2 are estimated directly and in Λ the path coefficients of the observed variables on their errors $(b_{e_i}^*)$ are estimated. This model has 10 distinct elements in Σ and 10 parameters to be estimated $(b_1^*, b_2^*, b_3^*, b_4^*, b_{e_1}^*, b_{e_2}^*, b_{e_3}^*, b_{e_4}^*, b_5^*, \rho_{e_1e_3})$, i.e., the model is just identified. The expected variance-covariance matrix $\Sigma = \Lambda\Phi\Lambda'$, i.e., the matrix Θ is taken to be zero.

The Figure 2.d. model poses two problems: the parameters b_3^*, b_4^*, and b_5^* are not identified and the expected correlation between X_4 and X_1 or X_2 is specified as zero even though the observed correlation may differ from zero presumably because of sampling fluctuations. The analysis in Section II showed that X_4 does not contribute to the identification of parameters, i.e., only the product $b_3^*b_5^*$ is identified with or without X_4. Without X_4 the model is that of Figure 1.b. and no purpose is served by retaining F_2. Assuming all variables are standardized $X_1 = b_1^*F_1 + \theta_1$ may be substituted for $F_1 = b_1^*X_1 + e_1$ as noted earlier. With F_2 eliminated and knowing that only the correlation of X_3 with F_1 is identified the model may be written as:

$$X_1 = b_1^*F_1 + \theta_1 \quad , \tag{10a}$$

$$X_2 = b_2^*F_1 + e_2 \quad , \text{ and} \tag{10b}$$

$$X_3 = b_3^*b_5^*F_1 + b_3^*b_4^*X_4 + e_3' \quad \text{where} \quad e_3' = b_3^*\theta + e_3 \quad . \tag{10c}$$

For convenience define $b_{35}^* = b_3^*b_5^*$ and $b_{34}^* = b_3^*b_4^*$. For computational simplicity define a new factor x_4 which is identical to the observed X_4, i.e., $X_4 = x_4$. The factors are then $\underset{\sim}{F}' = (F_1, x_4)$,

$$\Lambda = \begin{bmatrix} b_1^* & 0 \\ b_2^* & 0 \\ b_{35}^* & b_{34}^* \\ 0 & 1 \end{bmatrix} ,$$

$$\Phi = \begin{bmatrix} 1 & 0 \\ 0 & V_{x_4} \end{bmatrix} ,$$

and

$$\Theta^2 = \begin{bmatrix} V_{\theta_1} & 0 & 0 & 0 \\ 0 & V_{e_2} & 0 & 0 \\ 0 & 0 & V_{e'_3} & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} .$$

In the fourth row of Θ^2 the diagonal cell is zero to indicate the identity $X_4 = x_4$ without residuals. If the expected matrix Σ is computed, i.e., $\Sigma = \Lambda\Phi\Lambda' + \Theta^2$ we find:

$$\Sigma = \begin{bmatrix} V_{X_1} & b_1^*b_2^* & b_1^*b_{35}^* & 0 \\ b_1^*b_2^* & V_{X_2} & b_2^*b_{35}^* & 0 \\ b_1^*b_{35}^* & b_2^*b_{35}^* & V_{X_3} & b_{34}^* \\ 0 & 0 & b_{34}^* & V_{X_4} \end{bmatrix} .$$

This shows that the expected correlations of X_1 and X_2 with X_4 are zero. This follows from the specification in Φ that x_4 is uncorrelated with F_1.

In the analysis of the model in Figure 2.e., the correlations among X_1, X_2, and F_1 were identified first and then b_1^* and b_2^* identified from these correlations. The simplest estimation procedure is to estimate the correlations among X_1, X_2, and F_1 and then compute b_1^* and b_2^* from the estimated correlations. This problem can be handled by defining two factors $x_1 = X_1$ and $x_2 = X_2$. The structural equations are:

$$X_1 = x_1 \quad ,$$

$$X_2 = x_2 \quad ,$$

$$X_3 = b_3^* F_1 + e_3 \quad , \text{ and}$$

$$X_4 = b_4^* F_1 + e_4 \quad .$$

The factors are $\underset{\sim}{F}' = (x_1 , x_2 , F_1)$,

$$\Lambda = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & b_3^* \\ 0 & 0 & b_4^* \end{bmatrix} ,$$

$$\Phi = \begin{bmatrix} V_{x_1} & \rho_{x_1 x_2} & \rho_{x_1 F_1} \\ \rho_{x_1 x_2} & V_{x_2} & \rho_{x_2 F_1} \\ \rho_{x_1 F_1} & \rho_{x_2 F_1} & 1 \end{bmatrix} ,$$

and

$$\Theta^2 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & V_{e_3} & 0 \\ 0 & 0 & 0 & V_{e_4} \end{bmatrix} .$$

There are 10 distinct elements in Σ and nine parameters to be estimated ($b_3^*, b_4^*, V_{x_1}, V_{x_2}, \rho_{x_1 x_2}, \rho_{X_1 F_1}, \rho_{X_2 F_1}, V_{e_3}$, and V_{e_4}), so that the model has one overidentifying restriction. Note that the estimated elements of Φ should be used to estimate b_1^* and b_2^* ($r_{X_1X_2}$ may not equal $\rho_{x_1x_2}$) .

In relation to the model in Figure 2.f. Costner (1969) discussed the problem of ascertaining whether b_8^* was zero and of distinguishing the $b_8^* = 0$ model from one in which errors (e.g., e_3 and e_4) were correlated. To see how this is accomplished in Jöreskog's approach, first consider the model when $b_8^* = 0$ and treating residuals as factors:

$$\underset{\sim}{X}' = (X_1, X_2, X_3, X_4, X_5, X_6) \quad ,$$

$$\underset{\sim}{F}' = (F_1, F_2, e_1, e_2, e_3, e_4, e_5, e_6) \quad ,$$

$$\Lambda = \begin{bmatrix} b_1^* & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ b_2^* & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ b_3^* & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & b_4^* & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & b_5^* & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & b_6^* & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\Phi = \begin{bmatrix} 1 & b_7^* & 0 & 0 & 0 & 0 & 0 & 0 \\ b_7^* & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & V_{e_1} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & V_{e_2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & V_{e_3} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & V_{e_4} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & V_{e_5} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & V_{e_6} \end{bmatrix} ,$$

and

$$\Sigma = \Lambda\Phi\Lambda' \qquad (\text{i.e., } \Theta^2 = 0 \text{ }).$$

Note that we have chosen not to introduce the residual θ into the analysis because we wish to standardize both F_1 and F_2, in which case $\rho_{F_1F_2} = b_7^*$. This model is a variation of that in Figure 2.b. and all parameters are identified. There are 21 distinct elements in Σ and 14 parameters to be estimated so that there are seven overidentifying restrictions. To test $b_8^* \neq 0$, we specify $X_4 = b_8^*F_1 + b_4^*F_2 + e_4'$, i.e., in Λ the fourth row, first column element is left "free" instead of fixed = zero. This model has one more parameter to be estimated and therefore six overidentifying restrictions. Thus the original model is more restrictive and will therefore typically have a larger χ^2. In large samples, the difference in χ^2 between these two models, with degrees of freedom equal to the difference in number of restrictions, can be used to test the hypothesis that $b_8^* \neq 0$. Similarly the model with e_3 and e_4 correlated ("free") in Φ instead of independent (fixed = 0), would have six degrees of freedom and the difference in χ^2 with one degree of freedom would be a test of the hypothesis that e_3 and e_4 are uncorrelated. A comparison of the χ^2 for $b_8^* \neq 0$ to that for $\rho_{e_3e_4} \neq 0$ gives an indication of which is the better fitting model. Costner (1969, Figure 10) also raises the question of whether e_1 and e_2 are correlated. This hypothesis is tested by allowing the covariance between e_1 and e_2 in Φ to be "free," the change in χ^2 with one degree of freedom providing the appropriate statistical test. Hypotheses involving "constrained" parameters may be tested similarly, e.g., $b_1^* = b_6^*$ (Heise, 1969) or $V_{e_1} = V_{e_6}$ (Wiley & Wiley, 1970).

It can be observed that use of Jöreskog's program requires the investigator to know the identification status of each parameter, but does not require the complex algebraic manipulations provided by Costner (1969) and Blalock (1969). It is important to recognize the essentials of each model in order to fit it into Jöreskog's general model. Jöreskog's model assumes that the observed variables are "random" rather than "fixed" but it is doubtful that most applied sociologists need to be concerned about this issue which is minor in comparison to the usual questionable validity of measures and models.

Intraclass Reliability Estimates: Testing Structural Assumptions

Werts, C. E., Linn, R. L., and Jöreskog, K. G.

Abstract

Intraclass correlation reliability estimates are based on the assumption that the various measures are equivalent. Jöreskog's (1970) general model for the analysis of covariance structures can be used to test the validity of this assumption.

Intraclass Reliability Estimates: Testing Structural Assumptions[*]

Werts, C. E., Linn, R. L., and Jöreskog, K. G.

The validity of using intraclass correlation to estimate reliability is dependent on a variety of assumptions (Winer, 1962, Chapter 4; Cronbach, Rajaratnam, & Gleser, 1963; Stanley, 1971, pps. 420-429). This paper will focus on testing the assumption that the various measures are "equivalent" or "parallel" (Lord & Novick, 1968, pg. 48). Jöreskog's (1970) general model for the analysis of covariances structures will be used for this purpose. Some implications for generalizability theory (Cronbach, Rajaratnam, & Gleser, 1963; Rajaratnam, Cronbach, & Gleser, 1965; Gleser, Cronbach, & Rajaratnam, 1965) will be considered.

I. Jöreskog's General Model for the Analysis of Covariance Structures

Quoting Jöreskog, van Thillo, & Gruvaeus (1971, pg. 2-3):

"The general model considers a data matrix $X(N \times p)$ of N observations on p variates and assumes that the rows of X are independently distributed, each having a multivariate normal distribution with the same variance-covariance matrix Σ. It is assumed that

$$\varepsilon(X) = A\Xi P \quad , \qquad (1)$$

where $A(N \times g) = (a_{\alpha s})$ and $P(h \times p) = (p_{ti})$ are known matrices of ranks g and h, respectively, $g \leq N$, $h \leq p$ and $\Xi(g \times h) = (\xi_{st})$ is a matrix of parameters; and that Σ has the form

$$\Sigma = B(\Lambda\Phi\Lambda' + \Psi^2)B' + \Theta^2 \quad , \qquad (2)$$

[*] The research reported herein was performed pursuant to Grant No. OEG-2-7000 33(509) with the United States Department of Health, Education, and Welfare and the Office of Education.

where the matrices $B(p \times q) = (\beta_{ik})$, $\Lambda(q \times r) = (\lambda_{km})$, the symmetric matrix $\Phi(r \times r) = (\phi_{mn})$ and the diagonal matrices $\psi(q \times q) = (\delta_{kl}\Psi_k)$ and $\Theta(p \times p) = (\delta_{ij}\theta_i)$ are parameter matrices.

Thus the general model is one where means, variances and covariances are structured in terms of other sets of parameters that are to be estimated. In any application of this model, p, N and X will be given by the data, and g, h, q, r, A and P will be given by the particular application. In any such application we shall allow for any one of the parameters in Ξ, B, Λ, Φ, ψ and Θ to be known a priori and for one or more subsets of the remaining parameters to have identical but unknown values. Thus parameters 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.

The computer program estimates the free and constrained parameters of any such model by the maximum likelihood method and provides a test of goodness of fit of the whole model against the general alternative that P is square and Ξ and Σ are unconstrained. A test of a specified model (hypothesis) may be obtained, in large samples, by computing the maximum likelihood solution under the two models and then setting up the likelihood ratio test (see 1.5). In the special case when both Ξ and Σ are unconstrained, one may test a sequence of hypotheses of the form

$$C\Xi D = 0 \ , \qquad (3)$$

where $C(s \times g)$ and $D(h \times t)$ are given matrices of ranks s and t, respectively."

II. Application

For illustrative purposes consider the situation in which four alternate forms (ratings, etc.) of a test are administered to the same people; the testing conditions being such as to justify the assumption that the person's scores on the alternate forms are experimentally independent. In Cronbach's terminology the <u>facet</u> under consideration is alternate forms and there are four <u>conditions</u> of this facet under which each person is observed. The data would be analyzed with a two-way analysis of variance (ANOVA) model in which each row corresponds to the scores for a given person and each column to a different measure as shown in Table 1.

	Alternate Forms					
Person	X_1	X_2	X_3	X_4	Total	Mean
1	X_{11}	X_{12}	X_{13}	X_{14}	P_1	$\bar{P}_1$
2	X_{21}	X_{22}	X_{23}	X_{24}	P_2	$\bar{P}_2$
N	X_{N1}	X_{N2}	X_{N3}	X_{N4}	P_N	$\bar{P}_N$
Total	T_1	T_2	T_3	T_4	G	
Mean	$\bar{T}_1$	$\bar{T}_2$	$\bar{T}_3$	$\bar{T}_4$		$\bar{G}$

Table 1

From this table the mean squares between people (MS_b) , mean squares within people (MS_w) and residual mean squares (MS_r) can be computed as shown in Winer (1962, Chapter 4). Following Cronbach, et al., (1963), the reliability (ρ_i) of the ith measure and the reliability (ρ_c) of a composite measure may be estimated as (p = # of measures):

$$\hat{\rho}_i = \frac{MS_b - MS_r}{MS_b + (p - 1)\ MS_r} \quad \text{and} \qquad (4)$$

$$\hat{\rho}_c = \frac{MS_b - MS_r}{MS_b} \;. \qquad (5)$$

These formulae do not assume that the expected value of the test means are equal; however if the expected value (μ) of the test means is constant (i.e., observed mean differences due to sampling error) then it would be appropriate to use:

$$\hat{\rho}_i = \frac{MS_b - MS_w}{MS_b + (p - 1)MS_w} \quad \text{and} \qquad (6)$$

$$\hat{\rho}_c = \frac{MS_b - MS_w}{MS_b} \qquad (7)$$

To test assumptions using Jöreskog's method we can start with a model in which the test means are assumed to differ and all measures have the same underlying true score, $\mathcal{T}$. In terms of equations (1) and (2) this corresponds to a single factor ($\mathcal{T}$) model where the observed vector is

$$X' = (X_1 , X_2 , X_3 , X_4) ,$$

$$A = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} ,$$

$$\Xi = [\mu_1 , \mu_2 , \mu_3 , \mu_4] ,$$

$$\Lambda = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{bmatrix} ,$$

$$\Phi = [1] ,$$

$$\psi^2 = \begin{bmatrix} V_{e_1} & 0 & 0 & 0 \\ 0 & V_{e_2} & 0 & 0 \\ 0 & 0 & V_{e_3} & 0 \\ 0 & 0 & 0 & V_{e_4} \end{bmatrix} ,$$

Θ^2 is a null matrix and B an identity matrix.

In this formulation the means in Ξ, the factor loadings in Λ and the error variances in ψ^2 are free parameters to be estimated. For convenience the variance of the true scores V has been standardized (i.e., $V = 1$ in Φ). Since there are 10 distinct elements in Σ (i.e., $p(p + 1) \div 2$) and only eight parameters in Λ and ψ^2 to be estimated (i.e., b_1, b_2, b_3, b_4, V_{e_1}, V_{e_2}, V_{e_3}, and V_{e_4}), this model has two overidentifying restrictions (degrees of freedom). When the maximum likelihood estimation procedure is used, Jöreskog's program (Jöreskog, van Thillo, Gruvaeus, 1971) yields a chi square measure which, in large samples and assuming multivariate normality of observed variables, is a measure of the fit of the model to the data. In the illustration this χ^2 with 2 degrees of freedom may be used to test the assumption that the four measures have a common true score τ. If this hypothesis is rejected then the exact meaning of a reliability estimate is in doubt. Perhaps there is not a single underlying true factor and/or the error independence assumptions are violated. If the single factor model is not rejected then reliabilities may be obtained from parameter estimates, i.e.:

$$\hat{\rho}_i = \frac{\hat{b}_i^2}{\hat{b}_i^2 + \hat{V}_{e_i}} \quad \text{and} \qquad (8)$$

$$\hat{\rho}_c = \frac{\left(\sum_{i=1}^{p} \hat{b}_i\right)^2}{\left(\sum_{i=1}^{p} \hat{b}_i\right)^2 + \sum_{i=1}^{p} \hat{V}_{e_i}} \, . \qquad (9)$$

Given a minimum of three measures the b_i are identified given only the assumption of single factoredness. With two measures it is necessary to make additional assumptions (e.g., equal b_i) for identification.

The intraclass correlation and generalizability theory procedures assume that the measures all have the same units of measurement, i.e., are "essentially tau equivalent" (Lord & Novick, 1968, pg. 50). It would not be meaningful to average scores from measures with different units as is done in Table 1 to obtain person means. In Jöreskog's method equal units are equivalent to the assumption that that the regression weights b_i are equal, i.e., $b_1 = b_2 = b_3 = b_4 = b$ in Λ. Therefore the next step in the analysis with Jöreskog's program is to constrain the parameters in Λ to be equal, obtaining a new χ^2 estimate of the fit of model to the data. The χ^2 will have three additional degrees of freedom because of this constraint. The increase is χ^2 from the previous step (where single factoredness was tested) with three degrees of freedom, tests the hypothesis that the units of measurement are equal. If this hypothesis is rejected then the ANOVA formulation is rejected whether used for estimating reliability or for generalizability procedures. If the hypothesis of equal units is not rejected then the parameter estimates may be used to estimate reliability as follows (p = # measures):

$$\hat{\rho}_i = \frac{\hat{b}^2}{\hat{b}^2 + \hat{V}_{e_i}} \quad \text{and} \qquad (10)$$

$$\hat{\rho}_c = \frac{(p\hat{b})^2}{(p\hat{b})^2 + \sum_{i=1}^{p} \hat{V}_{e_i}} \qquad (11)$$

An exactly equivalent formulation is obtained if we fix all b_i in Λ equal to unity, allowing V_{τ} to be free, in which case $\hat{V}_{\tau}$ will replace $\hat{b}^2$ in equations (10) and (11).

The reliability of any single measure from equation (10) may vary because of differing error variances whereas equations (4) and (6) imply that all measures have the same reliability. It follows that it is necessary to test whether the error variances are indeed equal, i.e., $V_{e_i} = V_e$. The third step (in addition to previous constraints) in the analysis is to constrain the error variances in Ψ^2 to be equal, i.e., $V_{e_1} = V_{e_2} = V_{e_3} = V_{e_4} = V_e$. This will add three degrees of freedom and the increase in χ^2 from the second step tests the hypothesis of equal error variances. If this hypothesis is rejected then it may be asserted that equations (5) and (7) underestimate the composite reliability. If this hypothesis is not rejected then reliability estimates may be obtained directly from parameter estimates:

$$\hat{\rho}_i = \frac{\hat{b}^2}{\hat{b}^2 + \hat{V}_e} \qquad (12) \text{ and}$$

$$\hat{\rho}_c = \frac{(p\hat{b})^2}{(p\hat{b})^2 + p\hat{V}_e} \qquad . \qquad (13)$$

The estimates from equations (12) and (13) carry the same assumptions as equations (4) and (5) respectively, however different estimates may result because (12) and (13) are estimated under structural specifications which are assumed for (4) and (5), but not constrained to follow. Nonetheless equations (4) and (5) would in principle be appropriate in this situation.

If the expected variance-covariance matrix (2) is examined it will be seen that the expected variance (diagonal of Σ) for the different measures are equal as are the expected covariances between measures (off diagonal elements of Σ). This is precisely the configuration assumed in the ANOVA procedure when used for testing treatment (between measure) effects (Winer, 1963, pg. 124). Jöreskog's method may also be used to test these "treatment" effects, i.e., whether the test means differ. To do this we would make the additional constraints that the elements in Ξ be equal, i.e., $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu$. The resulting increase in χ^2 with three degrees of freedom can be used to test the hypothesis of equal means. If this hypothesis is rejected then equations (4) and (5) are more appropriate than (6) or (7). If the hypothesis is not rejected equations (12) and (13) would still be appropriate, however the parameter estimates will generally differ because of the restriction on the means.

Overall it may be observed that the above four analytical steps test the several aspects of the hypothesis that the different measures are "equivalent." If the hypotheses from each of the four steps are not rejected the implication is that observed differences in means, variances, and covariances between tests, are ascribable to sampling error.

III. Discussion

In essence, equation (9) estimates the reliability of a composite of the measures included in the study given the assumption of a common underlying true score. The χ^2 measure of fit associated with this specification is a test of the validity of this assumption. In contrast, the intraclass estimate of composite reliability assumes equivalent measures (implying a single true factor). From a structural perspective, the intraclass reliability estimate is therefore of limited applicability and even when measures are equivalent does not provide population estimates which necessarily are constrained to be consistent with this assumption. Furthermore, the intraclass estimate is inappropiate when errors of measurement are nonindependent, e.g., if the measures were ratings and a single judge did two of the ratings, the errors for these two measures would probably not be experimentally independent due to halo effects. In this situation a single factor would not account for the covariances among measures. Using Jöreskog's method a model could be used which would allow for the appropriate pair of errors to be correlated (Werts & Linn, in press). In this case, application of equation (9) would estimate the squared correlation of the composite score to the true score, whereas equation (5) would yield meaningless results. Given matched (all persons take all measures) data, certain aspects of generalizability theory may be considered in light of the model developed in section II. In particular, Cronbach, et al., (1963) require the investigator to specify a universe of conditions of observation over which he wishes to generalize. The example in section II corresponds to a single facet design and an investigator might for example specify conditions i = 1,2 as the universe appropriate to his particular study. In our approach, equation (8) would provide the reliability estimates for individual measures and in equation (9) sums would be taken over i = 1,2 to provide the composite reliability for this particular universe. If we wished to assume (perhaps because of a χ^2 test) that the measures have the same units of measurement (as does generalizability theory),then equations (10) and (11) would apply. Generalizability theory is clearly superior to intraclass correlation procedures in not requiring equivalent measures, but is not as flexible as Jöreskog's approach because of the equal units assumption. Cronbach, et al., (1963) indicate that the observed scores are determined by the person's universe (i.e., "true") score defined as the first centroid factors of the covariances between conditions in the universe, other centroid factors required to account for covariances between conditions, and residual variance after removal of the factors. The variance of the observed scores for a particular measure equals the squared factor loading on the universe score plus the sum of squared loading on the other centroid factors plus residual variance. From a structural perspective this formulation is problematical because: (a) The first factor may not be the factor of interest, e.g., "methods" factors (Campbell & Fiske, 1959) frequently account for larger proportions of observed variance than "true," "trait," or "universe" factors. (b) In <u>reality</u> there may be several underlying "true" factors and/or "other" factors, which may be oblique.

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

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. Theory of generalizability: a liberalization of reliability theory. British Journal of Statistical Psychology, 1963, 16, 137-163.

Gleser, G. C., Cronbach, J. J., & Rajaratnam, N. Generalizability of scores influenced by multiple sources of variance. Psychometrica, 1965, 30, 395-418.

Jöreskog, K. C., van Thillo, M., & Gruvaeus, G. T. ACOVSM - a general computer program for analysis of covariance structures including generalized MANOVA. Research Bulletin 71-1. Educational Testing Service, Princeton, N. J., January 1971.

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

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

Rajaratnam, N., Cronbach, L. J., & Gleser, G. C. Generalizability of stratified-parallel tests. Psychometrica, 1965, 30, 39-56.

Stanley, J. C. Reliability. In Thorndike, R. L., (Ed.) Educational and Psychological Measurement, American Council on Education, Washington, D. C., 1971.

Werts, C. E., & Linn, R. L. Corrections for attenuation. Educational and Psychological Measurement, in press.

Winer, B. J. Statistical principles in experimental design. New York, McGraw Hill, 1962.