Evaluating Structural Equation Models with Unobservable Variables and Measurement Error

Claes Fornell; David F. Larcker

Journal of Marketing Research, Vol. 18, No. 1 (Feb., 1981), 39-50.

CLAES FORNELL AND DAVID F. LARCKER*

The statistical tests used in the analysis of structural equation models with unobservable variables and measurement error are examined. A drawback of the commonly applied chi square test, in addition to the known problems related to sample size and power, is that it may indicate an increasing correspondence between the hypothesized model and the observed data as both the measurement properties and the relationship between constructs decline. Further, and contrary to common assertion, the risk of making a Type II error can be substantial even when the sample size is large. Moreover, the present testing methods are unable to assess a model's explanatory power. To overcome these problems, the authors develop and apply a testing system based on measures of shared variance within the structural model, measurement model, and overall model.

Evaluating Structural Equation Models with Unobservable Variables and Measurement Error

As a result of Jöreskog's breakthrough in developing a numerical method for the simultaneous maximization of several variable functions (Jöreskog 1966) and his formulation of a general model for the analysis of covariance structures (Jöreskog 1973), researchers have begun to realize the advantages of structural equation models with unobserved constructs for parameter estimation and hypothesis testing in causal models (Anderson, Engledow, and Becker 1979; Andrews and Crandall 1976; Hargens, Reskin, and Allison 1976; Long 1976; Rock et al. 1977; Werts et al. 1977). Marketing scholars have been introduced to this powerful tool of analysis by the work of Bagozzi (1977, 1978a, b, 1980), Bagozzi and Burnkrant (1979), and

Aaker and Bagozzi (1979). As demonstrated in these applications, an important strength of the method is its ability to bring together psychometric and econometric analyses in such a way that some of the best features of both can be exploited. It is possible to form econometric structural equation models that explicitly incorporate the psychometrician's notion of unobserved variables (constructs) and measurement error in the estimation procedure. In marketing applications, theoretical constructs are typically difficult to operationalize in terms of a single measure and measurement error is often unavoidable. Consequently, given an appropriate statistical testing method, the structural equation models are likely to become indispensable for theory evaluation in marketing. Unfortunately, the methods of evaluating the results obtained in structural equations with unobservables are less developed than the parameter estimation procedure. Both estimation and testing are necessary for inference and the evaluation of theory. Accordingly, the purpose of our article is twofold: (1) to show that the present testing methods have several limitations and can give misleading results and (2) to present a more comprehensive testing method which overcomes these problems.

PROBLEMS IN THE PRESENT TESTING METHODS

The present testing methods for evaluating structural equation models with unobservable variables are a t-test and a chi square test. The t-test (the ratio of the parameter estimate to its estimated standard error) indicates whether individual parameter estimates are statistically different from zero. The chi square statistic compares the "goodness of fit" between the covariance matrix for the observed data and covariance matrix derived from a theoretically specified structure (model).

Two problems are associated with the application of t-tests in structural equation models. First, the estimation procedure does not guarantee that the standard errors are reliable. Although the computation of standard errors by inverting the Fisher information matrix is less prone to produce unstable estimates than earlier methods, complete reliance on t-tests for hypothesis testing is not advisable (Lee and Jennrich 1979). Second, the t-statistic tests the hypothesis that a *single* parameter is equal to zero. The use of t-tests on parameters understates the overall Type I error rate and multiple comparison procedures must be used (Bielby and Hauser 1977).

The evaluation of structural equation models is more commonly based on a likelihood ratio test (Jöreskog 1969). Assume that the null hypothesis (H_0) is that the observed covariance matrix (S) corresponds to the covariance matrix derived from the theoretical specification (Σ) and that the alternative hypothesis (H_1) is that the observed covariance matrix is any positive definite matrix. For these hypotheses, minus twice the natural logarithm of the likelihood ratio simplifies to

(1) $\qquad N \cdot F_0 \sim \chi^2 \left[\dfrac{1}{2}(p + q)(p + q + 1) - Z \right]$

where:

 N is the sample size,

 F_0 is the minimum of fitting function $F = \log |\Sigma| + \mathrm{tr}(S\Sigma^{-1}) - \log |S| - (p + q)$,

 Z is the number of independent parameters estimated for the hypothesized model,

 q is the number of observed independent variables (x), and

 p is the number of observed dependent variables (y).

The null hypothesis (S = Σ) is rejected if $N \cdot F_0$ is greater than the critical value for the chi square at a selected significance level.

The first problem with the chi square test is that its power (ability to reject the null hypothesis, H_0, when it is false) is unknown (Bielby and Hauser 1977). Knowledge of the power curve of the chi square is critical for theory evaluation in structural equation

models because the testing is organized so that the researcher *a priori* expects that the null hypothesis will not be rejected. In contrast, in most significance testing the theory is supported if H_0 is rejected. If the power of the chi square test is low, the null hypothesis will seldom be rejected and the researcher using structural equation models may accept a false theory, thus making a Type II error.

The usual method of reducing the seriousness of Type II errors is to set up a sequence of hypothesis tests such that a model (A) is a special case of a less restrictive model (B). Assume that F_A and F_B are the minimums of equation 1 and Z_A and Z_B are the number of independent parameters estimated for models A and B, respectively. The appropriate statistic to test the null hypothesis that $S = \Sigma_A$ versus the alternative hypothesis that $S = \Sigma_B$ is

(2) $\qquad N \cdot (F_A - F_B) \sim \chi^2_{(Z_B - Z_A)}.$

This approach is consistent with traditional hypothesis testing in that model B is supported only if the null hypothesis is rejected, and thus the importance of Type II errors is reduced. This testing procedure, however, is applicable only to "nested models" (Aaker and Bagozzi 1979; Bagozzi 1980; Jöreskog and Sörbom 1978).[1] Furthermore, the sequential tests are not independent and the overall Type I error rate for the series of tests is larger than that of the individual tests.

The chi square test has two additional limitations which, to our knowledge, have not been considered in prior theoretical discussions or applications of structural equation models. The first and most serious of these is that the test may indicate a good fit between the hypothesized model and the observed data even though both the measures and the theory are inadequate. In fact, the goodness of fit can actually improve as properties of the measures and/or the relationships between the theoretical constructs decline. This problem has important implications for theory testing because it may lead to acceptance of a model in which there is no relationship between the theoretical constructs.

A second limitation of the chi square is related to the impact of sample size on the statistic (Jöreskog 1969). For example, if the sample size is small, $N \cdot F_0$ may not be chi square distributed. Further, the risk of making a Type II error usually is assumed to be reduced sharply with increasing sample size. However, contrary to this assumption, the risk of making a Type II error can be substantial even when the sample size is large.

---

[1]See Bentler and Bonett (1980) for other approaches to the assessment of goodness of fit.

SIMULATION ANALYSIS

To demonstrate the problems and show how they confound the evaluation of structural equation models, we performed two simulations. Before presenting the simulation results, we introduce the terminology and notation used. Let R_{xx}, R_{yy}, and R_{xy} refer to the $(q \times q)$ correlation matrix of x variables, the $(p \times p)$ correlation matrix of y variables, and the $(q \times p)$ correlation matrix between the x and y variables, respectively. For our purposes, the term *theory* refers to the statistical significance of the R_{xy} correlations. For example, if all R_{xy} correlations are statistically significant, the observed data have significant theory (the relationship between the independent and dependent variables is significant). If at least one of the R_{xy} correlations is statistically insignificant, the observed data are considered to have insignificant theory. The term *measurement* refers to the size and statistical significance of the R_{xx} and R_{yy} correlations. As these correlations become larger (smaller), the convergent validity or reliability of the associated x and y constructs becomes higher (lower). The term *discriminant validity* refers to the relationship of the off-diagonal terms of R_{xx} and R_{yy} with R_{xy}. Because the x variables and y variables are indicators of different constructs, discriminant validity is exhibited only if all the correlations in R_{xx} and R_{yy} (measurement) are statistically significant and each of these correlations is larger than all correlations in R_{xy} (Campbell and Fiske 1959).

The notation used in our analysis is introduced by expressing the $[(p + q) \times (p + q)]$ correlation matrix as a linear structural relations model. This model consists of two sets of equations: the structural equations and the measurement equations. The linear structural equation is

$$(3) \qquad B\eta = \Gamma\xi + \zeta$$

where:

B is an $(m \times m)$ coefficient matrix ($\beta_{ij} = 0$ means that η_j and η_i are not related),
Γ is an $(m \times n)$ coefficient matrix ($\gamma_{ij} = 0$ means that η_i is not related to ξ_j),
η is an $(m \times 1)$ column vector of constructs derived from the dependent variables (y),
ξ is an $(n \times 1)$ column vector of constructs derived from the independent variables (x),
ζ is an $(m \times 1)$ column vector of the errors in the structural equations,
m is the number of constructs (latent variables developed from the observed dependent variables, and
n is the number of constructs (latent variables) developed from the observed independent variables.

The measurement equations are

$$(4) \qquad y = \Lambda_y \eta + \epsilon$$

and

$$(5) \qquad x = \Lambda_x \xi + \delta$$

where:

y is a $(p \times 1)$ column vector of observed dependent variables,
x is a $(q \times 1)$ column vector of observed independent variables,
Λ_y is a $(p \times m)$ regression coefficient matrix of y on η,
Λ_x is a $(q \times n)$ regression coefficient matrix of x on ξ,
ϵ is a $(p \times 1)$ column vector of errors of measurement in y,
δ is a $(q \times 1)$ column vector of errors of measurement in x,
Φ is the $(n \times n)$ covariance matrix of ξ,
ψ is the $(m \times m)$ covariance matrix of ζ,
θ_ϵ is the $(p \times p)$ covariance matrix of ϵ, and
θ_δ is the $(q \times q)$ covariance matrix of δ.

To simplify our presentation, we conducted the simulation analysis by using a two-construct model in which each construct had two indicators. For this model, B is an identity matrix, θ_ϵ and θ_δ are diagonal matrices, Φ, ψ, and Γ have only a single element, and both Λ_y and Λ_x are a (2×1) column vector. The two-indicator/two-construct model is illustrated in Figure 1.

The properties of the chi square statistic can be observed by manipulating the theory and measurement of a single data set. The first correlation matrix in Table 1 (100% measurement, 100% theory) was used as the basis of the first simulation, the purpose of which was to examine the relationship between the chi square statistic and changes in measurement and theory correlations. This data set was chosen to represent a realistic, positive-definite matrix and to

Figure 1
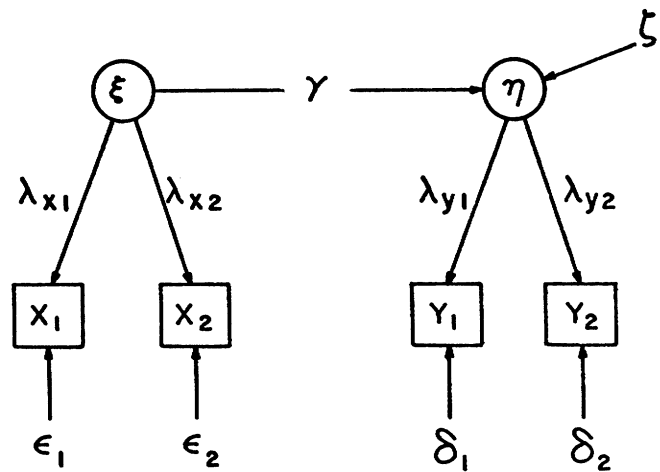### THE TWO-INDICATOR/TWO-CONSTRUCT MODEL

Table 1
SELECTED CORRELATION MATRICES FOR THE FIRST SIMULATION[a]

	100% measurement, 100% theory						100% measurement, 15% theory			
	y_1	y_2	x_1	x_2			y_1	y_2	x_1	x_2
y_1	1.0					y_1	1.0			
y_2	.633[b]	1.0				y_2	.633[b]	1.0		
x_1	−.355[b]	−.436[b]	1.0			x_1	−.137	−.170	1.0	
x_2	−.224[b]	−.414[b]	.413[b]	1.0		x_2	−.087	−.161	.413[b]	1.0

	64% measurement, 100% theory						64% measurement, 15% theory			
	y_1	y_2	x_1	x_2			y_1	y_2	x_1	x_2
y_1	1.0					y_1	1.0			
y_2	.510[b]	1.0				y_2	.510[b]	1.0		
x_1	−.355[b]	−.436[b]	1.0			x_1	−.137	−.170	1.0	
x_2	−.224[b]	−.414[b]	.330[b]	1.0		x_2	−.087	−.161	.330[b]	1.0

[a]A sample size of 200 was assumed.
[b]Statistically significant at the .01 level, two tail.

allow reductions in the correlations of R_{xx}, R_{yy}, and R_{xy} such that the resulting correlation matrices were (1) positive-definite, (2) internally consistent (i.e., to a certain extent, R_{xy} must be a function of R_{xx} and R_{yy}), and (3) large enough to be empirically identified in the maximum likelihood estimation (Kenny 1979). To be consistent with these constraints, the squared correlation for the off-diagonal elements of both R_{xx} and R_{yy} (i.e., measurement correlations) was arbitrarily decreased to 80% and 64% of the square of the correlations from the first correlation matrix. The squared correlations for all elements of R_{xy} were also arbitrarily decreased to 80%, 54%, and 15% of the square of the correlations shown in Table 1. Thus, 12 possible combinations of measurement and theory were evaluated in the first simulation. The extremes of the manipulation are presented in Table 1.

For purposes of illustration, a sample size of 200 was assumed and applied consistently in all simulations. The absolute level of significance obviously would differ for other sample sizes. However, the impact of sample size is not critical because our concern is not with absolute significance levels, but rather with how changes in measurement and theory correlations increase or decrease the chi square statistic. From equation 1 and for a specific S, it is clear that sample size can be viewed as a scaling factor and it does not affect F_0. Thus, the absolute size of the chi square statistic will change, but the direction of the change for different levels of theory and measurement correlation is unaffected by sample size.

The properties of the correlation matrices (significance of theory) and the chi square statistic from estimation of the structural equation model are presented in Table 2.[2] The four correlation matrices of Table 1 are the data used to calculate the results shown in the four corner cells of Table 2. These results

illustrate two problems with the chi square statistic. First, the last column of Table 2 suggests that a good fit (usually considered to be $p \geq .10$) can be obtained even if there is no statistically significant relationship between x and y variables (in this case, all elements of R_{xy} are statistically insignificant at the .01 level). Second, the chi square goodness of fit improves as both theory and measurement decrease!

To explain why goodness of fit improves as theory and/or measurement decline, let us examine what type of observed correlation matrix, S, will perfectly fit a two-construct/two-indicator model. The criterion for a perfect fit is structural consistency, which implies that all elements of R_{xy} are identical. If structural consistency is violated, goodness of fit will suffer. For example, if the correlations in R_{xy} differ widely, a two-construct model (involving a single γ coefficient to represent the relationship between x and y variables) will not adequately summarize the relationships between the original variables. This is because large divergence between the elements in R_{xy} suggests that there is more than one construct relationship between x and y variables. Thus, if the data are forced into an inappropriate two-construct structure, a poor fit will result.

The impact of structural consistency is demonstrated by estimating the correlation matrices in Table 3 (perfect structural consistency) and Table 4 (imperfect structural consistency) as a two-construct/two-indicator model. As expected, the chi square statistic is zero and the resulting fit between the observed data and the hypothesized structure is perfect when the correlation matrix has perfect structural consistency. However, if the structural consistency is reduced (Table 4), the chi square statistic indicates an extremely poor fit. Note that the difference between the correlation matrices in Tables 3 and 4 is the dispersion between the elements in R_{xy}. As dispersion increases, the difference between the observed correlation matrix, S, and the theoretical correlation matrix, Σ, for the two-indicator/two-construct model

---

[2]The numerical solutions for the structural equation models were obtained by using LISREL IV.

Table 2
CHI SQUARE RESULTS FOR THE FIRST SIMULATION[a]

| | | Increasing theory | | | |
		100%[b]	80%[b]	54%[b]	15%[b]
Increasing measurement	100%[b]	Significant theory[c] 2.8334[d] (.0923)[e]	Significant theory 2.2903 (.1302)	Significant theory 1.5485 (.2134)	Insignificant theory .4381 (.5080)
	80%[b]	Significant theory 2.2576 (.1330)	Significant theory 1.8248 (.1767)	Significant theory 1.2342 (.2666)	Insignificant theory .3493 (.5545)
	64%[b]	Significant theory 1.8827 (.1700)	Significant theory 1.5220 (.2173)	Significant theory 1.0296 (.3103)	Insignificant theory .2915 (.5893)

[a]A sample size of 200 was assumed.
[b]The percentage refers to the amount that the squared measure (or theory) correlations were reduced in relation to the correlation matrix in Table 1 (100% measurement, 100% theory).
[c]Significant theory means that all the correlations in R_{xy} are statistically significant at the .01 level.
[d]The chi square statistic.
[e]The probability of obtaining a chi square with 1 degree of freedom of this value or larger.

increases and the chi square statistic increases.

If the elements of R_{xy} are not identical, the size of the chi square statistic in the two-construct model is also affected by the dispersion between R_{xx} and R_{yy}. For the results in Table 2, as the measurement and/or theory correlations are reduced, the dispersion within these two sets of correlations is also reduced.[3] Therefore, as theory and/or measurement decrease, the goodness of fit between S and Σ for the two-indicator/two-construct model increases and the chi square statistic decreases. In fact, the chi square statistic in this model can be expressed as

(6) $\chi_i^2 = b_1 \sigma_{mi}^2 \sigma_{Ti}^2$

where:

χ_i^2 is the ith chi square value,
σ_{mi}^2 is the ith variance of the off-diagonal correlations from R_{xx} and R_{yy}, $\sigma_{mi}^2 > 0$,
σ_{Ti}^2 is the ith variance of the correlations in R_{xy}, $\sigma_{Ti}^2 > 0$, and
b_1 is the positive coefficient or scaling factor.

A second simulation, independent of the first, was conducted to illustrate equation 6 empirically. The first correlation matrix in Table 5 (100% measurement, 100% theory) was used as the starting point for the manipulation. The squared correlation for the off-diagonal elements of both R_{xx} and R_{yy} (measurement)

[3]For example, assume that the two measurement correlations are .625 and .640 and that each is reduced by $1/\sqrt{2}$ to .442 and .453, respectively. The standard deviation of the first pair of correlations is .0075 and the standard deviation of the second pair is .0055. Thus, as the correlations are decreased, the dispersion (in this case measured by the standard deviation) between correlations also decreases.

was decreased to 50%, 25%, 12.5%, and 6.25% of the square of the correlations shown in Table 5. The reductions were chosen arbitrarily to satisfy the same constraints (positive-definite, internally consistent, and empirically identified matrices) as in the first simulation.

To avoid confounding effects, we excluded from analysis correlation matrices that violate construct validity (no discriminant and/or convergent validity). The chi square statistic and its associated probability

Table 3
AN EXAMPLE OF A CORRELATION MATRIX WHICH PERFECTLY FITS THE TWO-INDICATOR/ TWO-CONSTRUCT MODEL

	Y_1	Y_2	X_1	X_2
Y_1	1.0			
Y_2	.500	1.0		
X_1	.250	.250	1.0	
X_2	.250	.250	.500	1.0

	Standardized estimate
λ_{y1}	.707
λ_{y2}	.707
λ_{x1}	.707
λ_{x2}	.707
γ	.500
ψ	.750
θ_ϵ	$\begin{bmatrix} .500 & \\ & .500 \end{bmatrix}$
θ_δ	$\begin{bmatrix} .500 & \\ & .500 \end{bmatrix}$

chi square statistic = .000
d.f. = 1
p = 1.000

Table 4
AN EXAMPLE OF A CORRELATION MATRIX WHICH INADEQUATELY FITS THE TWO-INDICATOR/ TWO-CONSTRUCT MODEL

	Y_1	Y_2	X_1	X_2
Y_1	1.0			
Y_2	.500	1.0		
X_1	.350	.250	1.0	
X_2	.250	.350	.500	1.0

	Standardized estimate
λ_{y1}	.707
λ_{y2}	.707
λ_{x1}	.707
λ_{x2}	.707
γ	.600
ψ	.640

$$\theta_\epsilon \quad \begin{bmatrix} .500 & \\ & .500 \end{bmatrix}$$

$$\theta_\delta \quad \begin{bmatrix} .500 & \\ & .500 \end{bmatrix}$$

chi square statistic = 8.1236
d.f. = 1
p = .0044

level for the remaining 18 combinations of theory and measurement are presented in Table 6. As in the first simulation, the observed goodness of fit improves as theory and/or measurement decline. If we regress the chi square values from Table 6 on the product of the variances (equation 6) and suppress the intercept, the regression coefficient, b_1, is equal to 72.73 × 10^6 and the R^2 is approximately equal to one (actually .997). The estimate for b_1 can now be used to compute any chi square value in Table 6 from the respective correlation matrices. Because the chi square is a function of the product of the variance of the measurement correlations and the variance of the theory correlations, the results show that changes in the theory model can be compensated for by changes in

the measurement model. Thus, an inadequate relationship between constructs (i.e., theory) may be partially offset by measurement properties, and using the chi square statistic for theory testing would be inappropriate.

Finally, we can easily demonstrate that the risk of making a Type II error, even in large samples, can be substantial. For example, the lower right cell of Table 2 is associated with a correlation matrix which has no statistically significant correlations in $\mathbf{R}_{xy}$, yet has a chi square value of .2915. The critical value for a chi square with one degree of freedom at the .01 level of statistical significance is 6.63. Using equation 1, we find that the sample size necessary to reject the theory is approximately 4550.

A PROPOSED TESTING SYSTEM

As applied to structural equations models, the chi square test has several limitations in addition to the well-known problems related to sample size sensitivity and lack of a defined power function. Further, the t-tests for individual parameter significance can be questioned on the basis of computational (Lee and Jennrich 1979) and probability/significance (Bielby and Hauser 1977) aspects. Therefore, a testing method is needed that:

—is sensitive to both convergent and discriminant validity in measurement,
—is sensitive to lack of significant theory,
—evaluates measurement and theory both individually and in combination so as to detect compensatory effects,
—is less sensitive to sample size,
—does not rely on standard errors from the estimation procedure, and
—is applied in such a way that a Type II error does not imply failure to reject a false theory.

By drawing on the statistical literature of canonical correlation analysis and by extending redundancy analysis to undefined factors, we have devised a testing

Table 5
SELECTED CORRELATION MATRICES FOR THE SECOND SIMULATION[a]

100% measurement, 100% theory						100% measurement, 6.25% theory			
	y_1	y_2	x_1	x_2		y_1	y_2	x_1	x_2
y_1	1.0				y_1	1.0			
y_2	.625[b]	1.0			y_2	.625[b]	1.0		
x_1	.327[b]	.367[b]	1.0		x_1	.081	.091	1.0	
x_2	.422[b]	.327[b]	.640[b]	1.0	x_2	.105	.081	.640[b]	1.0

6.25% measurement, 100% theory[c]						6.25% measurement, 6.25% theory			
	y_1	y_2	x_1	x_2		y_1	y_2	x_1	x_2
y_1	1.0				y_1	1.0			
y_2	.156[b]	1.0			y_2	.156[b]	1.0		
x_1	.327[b]	.367[b]	1.0		x_1	.081	.091	1.0	
x_2	.422[b]	.327[b]	.160[b]	1.0	x_2	.105	.081	.160[b]	1.0

[a] A sample size of 200 was assumed.
[b] Statistically significant at the .01 level, two tail.
[c] No construct validity.

Table 6
CHI SQUARE RESULTS FOR THE SECOND SIMULATION[a]

		100%[b]	50%[b]	Increasing theory 25%[b]	12.5%[b]	6.25%[b]
	100%[b]	6.4745[c] (.0109)[d]	3.1783 (.0746)	1.6196 (.2031)	.7710 (.3799)	.4035 (.5253)
Increasing measurement	50%[b]	2.8710 (.0902)	1.4159 (.2341)	.7235 (.3950)	.3448 (.5571)	.1806 (.6709)
	25%[b]	e	.9256 (.3360)	.4733 (.4915)	.2256 (.6348)	.1182 (.7310)
	12.5%[b]	e	e	.3677 (.5443)	.1753 (.6754)	.0918 (.7619)
	6.25%[b]	e	e	e	e	.0782 (.7798)

[a] A sample size of 200 was assumed.
[b] The percentage refers to the amount that the squared measure (or theory) correlations were reduced in relation to the correlation matrix in Table 5 (100% measurement, 100% theory).
[c] The chi square statistic.
[d] The probability of obtaining a chi square with 1 degree of freedom of this value or larger.
[e] No construct validity (no discriminant and/or convergent validity) was obtained for this correlation matrix.

system that satisfies these requirements. Unlike the current testing procedures, the proposed system is concerned with measures of explanatory power. We confine our discussion to the two-construct model (Figure 1) and assume that all variables and the constructs η and ξ are standardized. Generalizations are discussed after the derivation and application to the simulated data.

Evaluating the Structural Model

In canonical correlation analysis, several measures with associated statistical tests of construct relationships have been developed (Coxhead 1974; Cramer 1974; Cramer and Nicewonder 1979; Hotelling 1936; Miller 1975; Miller and Farr 1971; Roseboom 1965; Shaffer and Gillo 1974; Stewart and Love 1968). If η and ξ are standardized, γ^2 is equal to the squared correlation between the two constructs. Because the diagonal element of Ψ indicates the amount of unexplained variance, the shared variance in the structural model is

$$(7) \qquad \gamma^2 = 1 - \Psi.$$

The statistical significance of γ^2 can be assessed in several ways. If the measurement portion of the model is ignored and the constructs are treated as variables, the significance tests for γ^2 are analogous to the significance tests for the multiple correlation coefficient (R^2) in regression. Thus, the significance of γ^2 can be evaluated by the traditional F-test with 1 and $N - 2$ degrees of freedom.

Evaluating the Measurement Model

Before testing for a significant relationship in the structural model, one must demonstrate that the measurement model has a satisfactory level of validity

and reliability. From classical test theory, the reliability (ρ_{yT}) of a single measurement y is

$$(8) \qquad \rho_{(yT)} = \frac{\text{Var}(T)}{\text{Var}(T) + \text{Var}(\varepsilon)} = 1 - \frac{\text{Var}(\varepsilon)}{\text{Var}(y)}$$

where T is the underlying true score and ε is the error of measurement. If all variables have zero expectation, the true score is independent of measurement error and individual measurement errors are independent ($\mathbf{T\varepsilon'} = \mathbf{I}$ and $\mathbf{\varepsilon\varepsilon'} = \mathbf{I}$); the reliability (convergent validity) of each measure y in a single factor model can be shown to be[4]

$$(9) \qquad \rho_y = \frac{\lambda_y^2}{\lambda_y^2 + \text{Var}(\varepsilon_y)}.$$

The reliability for the construct η is

$$(10) \qquad \rho_\eta = \frac{\left(\sum_{i=1}^{p} \lambda_{yi}\right)^2}{\left(\sum_{i=1}^{p} \lambda_{yi}\right)^2 + \sum_{i=1}^{p} \text{Var}(\varepsilon_i)}.$$

To examine more fully the shared variance in the measurement model, it is useful to extend ρ_y to include several measurements. Though ρ_y indicates the reliability of a single measure and ρ_η the reliability of the construct, neither one measures the amount of variance that is captured by the construct in relation to the amount of variance due to measurement error. The average variance extracted $\rho_{vc(\eta)}$ provides this information and can be calculated as

---

[4] See Werts, Linn, and Jöreskog (1974) or Bagozzi (1980) for the derivation.

$$(11) \qquad \rho_{vc(\eta)} = \frac{\sum\limits_{i=1}^{p} \lambda_{yi}^2}{\sum\limits_{i=1}^{p} \lambda_{yi}^2 + \sum\limits_{i=1}^{p} \text{Var}(\varepsilon_i)}.$$

If $\rho_{vc(\eta)}$ is less than .50, the variance due to measurement error is larger than the variance captured by the construct η, and the validity of the individual indicators (y_i), as well as the construct (η), is questionable. Note that $\rho_{vc(\eta)}$ is a more conservative measure than ρ_η. On the basis of ρ_η alone, the researcher may conclude that the convergent validity of the construct is adequate, even though more than 50% of the variance is due to error. The reliability for each measure x (ρ_x), for the construct ξ (ρ_ξ), and for the average variance extracted $(\rho_{vc(\xi)})$ can be developed in an analogous manner.

In addition, $\rho_{vc(\cdot)}$ can be used to evaluate *discriminant* validity. To fully satisfy the requirements for discriminant validity, $\rho_{vc(\eta)} > \gamma^2$ and $\rho_{vc(\xi)} > \gamma^2$. Unlike the Campbell and Fiske approach, this criterion is associated with model parameters and recognizes that measurement error can vary in magnitude across a set of methods (i.e., indicators of the constructs).

Evaluating the Overall Model

To assess the significance and the explanatory power of the overall model, one must consider *both* measurement and theory. The squared gamma measures the variance shared by the constructs η and ξ, and thus indicates the degree of explanatory power in the structural model. If the measures were perfect and if the ξ construct were a perfect linear combination of the x-variables, γ^2 would be the explained variance of the overall model. However, given that measurement error is present, any combined summary statistic for the overall model must represent the intersection between measurement variance and theory variance.

(variance in measurement − error in measurement) ∩
(variance in theory − error in theory)

The error in theory (structural equation) is indicated by ψ. The error in measurement is composed of two parts. The first is related to the variance lost in the extraction of constructs. As discussed previously, this is equal to

$$(12) \qquad 1 - \rho_{vc(\eta)}.$$

The mean variance of the set of y variables that is explained by the *construct* of the other set (ξ) is the redundancy between the sets of variables (Fornell 1979; Stewart and Love 1968; Van den Wollenberg 1977). For the two-construct model, the redundancy is

$$(13) \qquad \bar{R}_{y/\xi}^2 = \rho_{vc(\eta)}\gamma^2.$$

The second part of the error in measurement is related to the fact that, in contrast to canonical correlation analysis, redundancy $(\bar{R}_{y/\xi}^2)$ is *not* equivalent to average ability of the x *variables* to predict the y variables. The reason is that the constructs are undefined factors and *not* perfect linear combinations of their respective indicators. To develop a measure that allows operational inferences about variance shared by the observed variables (the model's explanatory power), one must determine the proportion of the ξ variance that is accounted for by a combination of its indicators. This second part of the measurement error can be obtained by regressing the ξ construct on the x variables, calculating the multiple correlation coefficient, and subtracting this value from one, or

$$(14) \qquad e_\xi = 1 - (\Lambda_x' \mathbf{R}_{xx}^{-1} \Lambda_x).$$

The resulting formula for explaining the y variables from the x variables (operational variance) becomes[5]

$$(15) \qquad \hat{\bar{R}}_{y/\xi}^2 = \rho_{vc(\eta)}\gamma^2(1 - e_\xi) = \bar{R}_{y/\xi}^2(1 - e_\xi).$$

The statistical significance of redundancy $(\bar{R}_{y/\xi}^2)$ and operational variance $(\hat{\bar{R}}_{y/\xi}^2)$ can be determined by an analogue to the traditional F-test in regression developed by Miller (1975). The test is

$$(16) \qquad \frac{\hat{\bar{R}}_{y/\xi}^2}{1 - \hat{\bar{R}}_{y/\xi}^2} \cdot \frac{(N - q - 1)}{q} \sim F_{q,(N-q-1)}.$$

Applying the Testing System to the Simulated Data

The proposed measures and their associated statistical tests were applied to the data from simulation 1. The evaluation of the structural model by use of gamma squared and an F-test is presented in Table 7. The four correlation matrices of Table 1 are the data used to calculate the results shown in the four corner cells of Table 7. Contrary to the results of the chi square in Table 2, γ^2 decreases (indicating a weaker relationship between constructs η and ξ) as the theory correlations are decreased. Because the shared variance between constructs decreases as the theory correlations are decreased, γ^2 behaves as expected. However, γ^2 increases as the measurement is reduced. The reason is that as measurement correlations are reduced, the dispersion in the correlation matrix becomes smaller (has more structural consistency), and the observed data have a better goodness of fit for the two-construct model in the sense that more variance is captured. In fact, the increase in γ^2 is large enough to indicate that the relationship between constructs in the right column is statistically significant, even though there are no statistically

---

[5] Both redundancy and operational variance are nonsymmetrical. If one is interested in explaining the variance in the x variables rather than the y variables, analogous measures can be easily developed for equations 12, 13, 14, and 15.

Table 7
EVALUATION OF THE STRUCTURAL MODEL FOR SIMULATION 1[a]

| | | | Increasing theory | | |
		100%[b]	80%[b]	54%[b]	15%[b]
Increasing measurement	100%[b]	.438[c] (76.8)[d]	.351 (53.3)	.233 (29.9)	.067 (7.1)
	80%[b]	.549 (119.9)	.439 (77.1)	.292 (41.2)	.083 (8.9)
	64%[b]	.686 (215.2)	.549 (119.9)	.365 (56.6)	.104 (11.4)

[a] A sample size of 200 was assumed.
[b] The percentage refers to the amount that the squared measure (or theory) correlations were reduced in relation to the correlation matrix in Table 1 (100% measurement, 100% theory). All values in parentheses are statistically significant at the .01 level (critical $F_{1,198} = 6.76$).
[c] Gamma squared (γ^2).
[d] F-test.

significant correlations between variables.[6]

The evaluation of the measurement models for the η and ξ constructs using the reliability (ρ) and average variance extracted (ρ_{vc}) is presented in Table 8. As expected, both statistics decrease as the measurement correlations are reduced and are unaffected by reductions in the theory correlations. For the ξ construct, we observe a problem with the traditional measure of reliability (ρ_ξ). For example, in the first row of Table 8, the calculated reliability may be considered acceptable ($\rho > .58$), although almost 60% ($1 - .414$)

---

[6] The same phenomenon occurs in canonical correlation analysis because the two-construct model discussed here is a reduced rank model. As structural consistency improves, more of the variance in R_{xy} is captured in a reduced rank (Fornell 1979).

of the variance, as indicated by $\rho_{vc(\xi)}$, is due to error. Though the interpretation of these statistics is subjective, the results cause us to question the convergent validity for the indicators of the ξ construct.

The evaluation of the overall model using operational variance ($\hat{R}^2_{y/\xi}$) and Miller's significance test is presented in Table 9. As expected, and contrary to the chi square results in Table 2, redundancy decreases as theory is reduced. We also find that operational variance for the right column is *not* statistically significant. This finding is consistent with the knowledge that there are no statistically significant correlations between the x and y variables. In a manner similar to the behavior of γ^2, operational variance increases as measurement is reduced. The explanation for this result is that as measurement decreases, the relationship between constructs (γ) increases more than the average variance extracted ($\rho_{vc(\xi)}$) decreases.

As we have shown, the measures associated with the new testing system are able to overcome the problems inherent in the present testing methods. The average variance extracted is sensitive to a lack of convergent validity and can be used to assess discriminant validity. The operational variance measure and its associated statistical test are sensitive to reductions in correlations between the x and y variables. Although this measure is a product of both structural model and measurement model results, these inputs can be evaluated separately so as to detect compensatory effects. Because the statistical tests are based on the F-distribution, they are somewhat less sensitive to sample size than a test based on the chi square distribution. None of the proposed tests utilize the standard errors from the estimation procedure. Finally, if a Type II error is made under this testing system, a false theory is not accepted.

Table 8
EVALUATION OF THE MEASUREMENT MODEL FOR SIMULATION 1[a]

		100%[b]		Increasing theory 80%[b]		54%[b]		15%[b]	
		η	ξ	η	ξ	η	ξ	η	ξ
Increasing measurement	100%[b]	.719[c] [.830][e]	.414[d] [.584][f]	.719 [.830]	.414 [.584]	.719 [.830]	.414 [.584]	.719 [.830]	.414 [.584]
	80%[b]	.642 [.775]	.370 [.539]	.642 [.775]	.370 [.539]	.642 [.775]	.370 [.539]	.642 [.775]	.370 [.539]
	64%[b]	.574 [.722]	.331 [.479]	.574 [.722]	.331 [.479]	.574 [.722]	.331 [.479]	.574 [.722]	.331 [.479]

[a] A sample size of 200 is assumed.
[b] The percentage refers to the amount that the squared measure (or theory) correlations were reduced in relation to the correlation matrix in Table 1 (100% measurement, 100% theory).
[c] Average variance extracted for η construct ($\rho_{vc(\eta)}$).
[d] Average variance extracted for ξ construct ($\rho_{vc(\xi)}$).
[e] Reliability for η construct (ρ_η).
[f] Reliability for ξ construct (ρ_ξ).

Table 9
EVALUATION OF THE OVERALL MODEL FOR SIMULATION 1 [a]

		100%	Increasing theory 80% [b]	54% [b]	15% [b]
Increasing measurement	100% [b]	.185 [c] (22.3) [d,e]	.144 (16.6) [e]	.092 (9.9) [e]	.029 (2.94)
	80% [b]	.190 (23.1) [e]	.152 (17.6) [e]	.103 (11.3) [e]	.029 (2.94)
	64% [b]	.194 (23.7) [e]	.154 (18.8) [e]	.104 (11.4) [e]	.030 (3.05)

[a] A sample size of 200 was assumed.
[b] The percentage refers to the amount that the squared measure (or theory) correlations were reduced in relation to the correlation matrix in Table 1 (100% measurement, 100% theory).
[c] Operational variance for the overall model ($\hat{R}^2_{y/\xi}$).
[d] Miller's statistic.
[e] Statistically significant at the .01 level (critical $F_{2,197} = 4.71$).

Limitations

Although we have shown that the proposed evaluation system is a useful extension and improvement to the testing procedures currently used, it is not free of limitations. Because the system builds on the linear model, the typical assumptions of multiple linear regression (independent and normally distributed error terms with a mean of zero and a constant variance, etc.) are required.

All significance testing procedures are sensitive to sample size and the F-test is no exception. For a sufficiently large sample, the F-test will indicate that the level of shared variance is different from zero. At that point, the issue becomes one of assessing the operational significance of the results. Unlike the present testing system which provides only a significance level for the goodness of fit, the proposed system can assess significance levels and indicate the size of the shared variance. If the shared variance is not large enough to warrant interpretation in terms of operational significance, one may of course reject the model regardless of its statistical significance.

Another concern is that all the measures proposed are summary statistics. As such, they cannot capture the full complexity of multivariate relationships. Instead, they attempt to simplify these relationships and provide indices for evaluating empirical models. The measure of operational variance for the overall model is the least sensitive to variations in individual parameter values. It is a composite measure that is also compensatory. In fact, the simulation analysis showed that several of the proposed measures (gamma squared, redundancy, and operational variance) indicated increasing significance when measurement correlations were reduced. Clearly, the reduction in the average variance extracted ($\rho_{vc(\eta)}$) was compensated

for in the overall indices. However, as pointed out before, the detection of compensatory effects is not complicated.

GENERALIZATIONS

The measures and statistical tests proposed are not restricted to the two-construct model. For simplicity, assume that η and ξ are standardized. The shared variance in the structural model (between constructs) can be calculated by generalizing equation 7. The appropriate measure is the redundancy *between constructs* which is identical to the average squared multiple correlation coefficient between each construct in η and all constructs in ξ,

$$(17) \qquad \bar{R}^2_{\eta/\xi} = \frac{1}{m} \mathrm{tr} \ [R^{-1}_{\xi\xi} R_{\xi\eta} R_{\eta\xi}],$$

where $R_{\xi\xi}$ is the $(n \times n)$ matrix of correlations between the constructs in ξ and $R_{\xi\eta}$ is the $(n \times m)$ matrix of correlations between ξ and η. The statistical significance of $\bar{R}^2_{\eta/\xi}$ can be determined by using Miller's test from equation 16 with $(n \cdot m)$ and $(N - n - 1) \cdot m$ degrees of freedom.

The measurement model tests are easily extended to a multiple construct model. No additional assumptions about construct independence, etc., are necessary. Therefore, equation 11 is directly generalizable to more complex models.

If the overall model is specified as a canonical model (Miller and Farr 1975), redundancy and operational variance are directly generalizable. These statistics are simply the sum of the respective measures for all two-construct pairs in the overall model. However, for more complex and less restrictive models with a multitude of construct associations, a single measure such as redundancy or operational variance cannot provide satisfactory information about the multivariate relationships involved. In such cases, the model can be partitioned into subsets of two-construct systems and a separate measure of redundancy (equation 13) and operational variance (equation 15) can be obtained for each subset.

The overall evaluation of a complex model for operational variance can also be accomplished by examining the *upper bound* of explanatory power, given the observed variables in the model. The trace of the matrix product $R^{-1}_{xx} R_{xy} R_{yx}$ is the total variance of the criterion variables that can be explained by the predictor variables. By dividing this product by the number of criterion variables one can obtain an index that goes from zero to one for the upper bound of average predictability (Van den Wollenberg 1977). The statistical significance of the upper bound, which in our terminology would be the maximum level of operational variance, can be assessed by using Miller's test as in equation 16.

SUMMARY

Because of their flexibility, ability to bring psychometric and econometric theory together in a unified manner, and their continual methodological advancement (Bentler 1976; Bentler and Bonett 1979; Bentler and Weeks 1979; Jennrich and Clarkson 1980), structural equation models incorporating unobservable variables and measurement error will undoubtedly have increasing application in theory testing and empirical model building in marketing. A primary interpretative index for these models is a chi square statistic measuring the goodness of fit between the hypothesized structure and the observed data. However, as we demonstrate, this statistic has several serious limitations. To overcome these problems, we propose a more comprehensive testing system. Unlike the chi square statistic, this system is based on measures of explanatory power (shared variance) for the (1) structural model, (2) measurement model, and (3) overall model. These measures and their associated statistical tests provide an assessment of both the statistical and operational (practical) significance for the theory being tested. It is important to realize that this system is not a substitute for the chi square test. Instead, it complements the currently used testing methods.

The simplest way to describe how the proposed testing system can be integrated with present testing methods is to outline the stages in the evaluation of theory via structural equation models. The first step is to determine whether the measures have satisfactory psychometric properties. The properties of interest are reliability (convergent validity), average variance extracted, and discriminant validity for each unobserved variable. The measurement model tests can be used to calculate the average variance extracted and provide a procedure, complementary to the traditional Campbell and Fiske approach, for establishing discriminant validity. (Obviously, if the measurement properties prove to be inadequate, it is not appropriate to proceed to theory testing.)

The next step is to examine the chi square value and determine its statistical significance. If the probability of obtaining a chi square larger than the observed chi square is less than some selected critical value (usually .10), the hypothesized model structure is rejected.[7] If the probability is greater than the selected critical value, the researcher may conclude that the data fit the hypothesized structure.[8] However, as we

demonstrate, a good fit is not a sufficient criterion for concluding that one's theory is supported by the data.

The manner in which the theory evaluation is completed depends on the purpose of the research. If the purpose is theory testing without regard to the explanatory power of the model, focus should center on the relationships between unobservable constructs. Support for the theory requires that γ^2 be statistically significant. If the purpose is to explain or predict (via the hypothesized model) the variance of the observed y variables from the *unobserved x constructs* (ξ), the redundancy, $\bar{R}^2_{y/\xi}$, must be statistically significant. Finally, if the purpose is to obtain a measure indicating how well the x *variables* account (via the hypothesized model) for the variance in the y variables, operational variance, $\hat{\bar{R}}^2_{y/\xi}$ is the relevant statistic. For decision sciences such as marketing, it is often desirable for models to have high explanatory or predictive power (operational variance). Though all the proposed measures build on the notion of explanatory power via shared variance, only $\hat{\bar{R}}^2_{y/\xi}$ is a direct summary measure of operational variance.

REFERENCES

Aaker, David and Richard P. Bagozzi (1979), "Unobservable Variables in Structural Equation Models with an Application in Industrial Selling," *Journal of Marketing Research*, 16 (May), 147–58.

Anderson, Ronald D., Jack L. Engledow, and Helmut Becker (1979), "Evaluating the Relationships Among Attitudes Toward Business, Product Satisfaction, Experience, and Search Effort," *Journal of Marketing Research*, 16 (August) 394–400.

Andrews, Frank M. and Rick Crandall (1976), "The Validity of Measures of Self-Reported Well-Being," *Social Indicators Research*, 3, 1–19.

Bagozzi, Richard P. (1977), "Structural Equation Models in Experimental Research," *Journal of Marketing Research*, 14 (May), 209–26.

——— (1978a), "Reliability Assessment by Analysis of Covariance Structures," in *Research Frontiers in Marketing: Dialogues and Directions*, S. C. Jain, ed. Chicago: American Marketing Association, 71–5.

——— (1978b), "The Construct Validity of Affective, Behavioral, and Cognitive Components of Attitude by Analysis of Covariance Structures," *Multivariate Behavioral Research*, 13, 9–31.

——— (1980), *Causal Models in Marketing*. New York: John Wiley & Sons, Inc.

——— and Robert E. Burnkrant (1979), "Attitude Measurement and Behavior Change: A Reconsideration of Attitude Organization and its Relationship to Behavior," in *Advances in Consumer Behavior*, Vol. 6, William L. Wilkie,

---

[7]See Bagozzi (1980, p. 204–14) and Bentler and Bonnet (1980) for a discussion of how to diagnose problems with the structural or measurement models when the chi square statistic is large.

[8]The residuals (S − Σ), which illustrate the differences between the observed covariance matrix and the theoretical covariance matrix, can also be examined to determine whether the hypothesized structure is tenable. The advantage of residual analysis (checking to make sure that all residuals are less than, say, .05) is that the

---

method is independent of sample size. However, both Jöreskog (1969) and Long (1976) argue that the chi square test and calculation of confidence intervals for all free parameters are better ways to detect misspecifications.

ed. Ann Arbor, Michigan: Association for Consumer Research.

Bentler, P. M. (1976), "A Multistructure Statistical Model Applied to Factor Analysis," *Multivariate Behavioral Research*, 11, 3–25.

—— and D. G. Bonnett (1980), "Significance Tests and Goodness of Fit in Analysis of Covariance Structures," presented at annual meeting of The American Psychological Association, Div. 23, Montreal, Quebec, September 1–5.

—— and D. G. Weeks (1979), "Interrelations Among Models for the Analysis of Moment Structures," *Multivariate Behavioral Research*, 14, 169–85.

Bielby, William T. and Robert M. Hauser (1977), "Structural Equation Models," *Annual Review of Sociology*, 3, 137–61.

Campbell, Donald T. and Donald W. Fiske (1959), "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix," *Psychological Bulletin*, 56 (March), 81–105.

Coxhead, P. (1974), "Measuring the Relationship Between Two Sets of Variables," *British Journal of Mathematical and Statistical Psychology* 27, 205–12.

Cramer, E. M. (1974), "A Generalization of Vector Correlations and its Relation to Canonical Correlations," *Multivariate Behavioral Research*, 9, 347–52.

—— and W. A. Nicewander (1979), "Some Symmetric, Invariant Measures of Multivariate Association," *Psychometrika*, 44 (March), 43–54.

Fornell, Claes (1979), "External Single-Set Components Analysis of Multiple Criterion/Multiple Predictor Variables," *Multivariate Behavioral Research*, 14, 323–38.

Hargens, L. L., B. Reskin, and P. D. Allison (1978), "Problems in Estimating Measurement Error from Panel Data: An Example Involving the Measurement of Scientific Productivity," *Sociological Methods & Research*, 4, 439–58.

Hotelling, H. (1936), "Relations Between Two Sets of Variables," *Biometrika*, 28, 321–77.

Jennrich, R. I. and D. B. Clarkson (1980), "A Feasible Method for Standard Errors of Estimate in Maximum Likelihood Factor Analysis," *Psychometrika*, 45, 237–47.

Jöreskog, Karl G. (1966), "UMFLA: A Computer Program for Unrestricted Maximum Likelihood Factor Analysis," Research Memorandum 66-20. Princeton, New Jersey: Educational Testing Service.

—— (1969), "A General Approach to Confirmatory Maximum Likelihood Factor Analysis," *Psychometrika*, 34 (June), 183–202.

—— (1973), "A General Method for Estimating a Linear Structural Equation System," in *Structural Equation Models in the Social Sciences*, A. S. Goldberger and O. D. Duncan, eds. New York: Seminar Press, 85–112.

—— and Dag Sörbom (1978), *LISREL IV: Analysis of Linear Structural Relationships by the Method of Maximum Likelihood*. Chicago: National Educational Resources, Inc.

Kenny, D. A. (1979), *Correlation and Causality*. New York: John Wiley & Sons, Inc.

Lee, S. Y. and R. I. Jennrich (1979), "A Study of Logarithms for Covariance Structure Analysis with Specific Comparisons Using Factor Analysis," *Psychometrika*, 4 (March), 99–113.

Long, J. Scott (1976), "Estimation and Hypothesis Testing in Linear Models Containing Measurement Error: A Review of Jöreskog's Model for the Analysis of Covariance Structures," *Sociological Methods & Research*, 5 (November), 157–206.

Miller, John K. (1975), "The Sampling Distribution and a Test for the Significance of the Bimultivariate Redundancy Statistic: A Monte Carlo Study," *Multivariate Behavioral Research* (April), 233–44.

—— and S. D. Farr (1971), "Bimultivariate Redundancy: A Comprehensive Measure of Interbattery Relationship," *Multivariate Behavioral Research*, 6, 313–24.

Rock, D. A., C. E. Werts, R. L. Linn, and K. G. Jöreskog (1977), "A Maximum Likelihood Solution to the Errors in Variables and Errors in Equations Model," *Multivariate Behavioral Research*, 12, 187–97.

Roseboom, W. W. (1965), "Linear Correlation Between Sets of Variables," *Psychometrika*, 30, 57–71.

Shaffer, J. P. and M. W. Gillo (1974), "A Multivariate Extension of the Correlation Ratio," *Educational and Psychological Measurement*, 34, 521–24.

Stewart, D. and W. Love (1968), "A General Canonical Correlation Index," *Psychological Bulletin*, 70, 160–3.

Van den Wollenberg, A. L. (1977), "Redundancy Analysis—An Alternative for Canonical Correlation Analysis," *Psychometrika*, 42 (June), 207–19.

Werts, C. E., R. L. Linn, and K. G. Jöreskog (1974), "Interclass Reliability Estimates: Testing Structural Assumptions," *Educational and Psychological Measurement*, 34, 25–33.

——, D. A. Rock, R. L. Linn, and K. G. Jöreskog (1977), "Validating Psychometric Assumptions Within and Between Populations," *Educational and Psychological Measurement*, 37, 863–71.