

***A Comparison of First and Second Generation Multivariate Analyses:
Canonical Correlation Analysis and Structural Equation Modeling*¹**

A. J. Guarino
Auburn University

Abstract

This study illustrates that Structural Equation Modeling (SEM) provides a more accurate representation of the latent variables as assessed by the structure coefficients than Canonical Correlation Analysis (CCA). A heuristic data set of five independent variables representing the Piagetian tasks measured by The Inventory of Piaget's Developmental Tasks (IPDT) and three dependent variables assessed by the Alabama Basic Competency Tests (BCT) were analyzed with both statistical methods. Although the relationship between IPDT and BCT was invariant, the corresponding latent variable structures differed for the two methods.

Educational research may be grossly divided into exploratory and confirmatory approaches. Exploratory analysis is used as a theory-generating procedure, whereas confirmatory procedures are employed as theory-testing procedures (Stevens, 2001). Statistical procedures that analyze exploratory research are known as "first generation" while statistical analyses designed to confirm a theory are known as "second generation" (Fornell, 1987). The decision as to which approach to use is contingent in part on the current understanding of the phenomenon being

¹ The author wishes to express his sincere appreciation and gratitude to Jeffrey Kromrey, Glenn Gamst, Larry Myers, and James Witte for their detailed review and scholarly suggestions.

investigated. A few examples of exploratory statistics with their confirmatory counterpart are (a) for univariate analysis (i.e., a single dependent variable), stepwise regression (exploratory) and hierarchical regression or path analysis (confirmatory); (b) for multivariate (i.e., multiple dependent variables), canonical correlation analysis (CCA) (exploratory) and structural equation modeling (SEM) (confirmatory). This dichotomy of exploratory and confirmatory research is obviously simplistic because much of the theoretical knowledge in the field of education is still in the formative stages (Chin, Marcolin, & Newsted, 1996). The use of the proper statistical analysis is often in doubt due to the difficulty of deciding when a theory is sufficiently developed to warrant a confirmatory analysis.

Advantages of Multiple Variables

Because of the complexity of the constructs in educational research, multidimensional statistical designs are now more prevalent. Multivariate methods have gained greater support in the research literature because “(multivariate methods) best honor the reality to which the researcher is purportedly trying to generalize” (Thompson, 1991, p. 80). Essential in comprehending multivariate methods is the understanding of the variate. The variate unfortunately has many names: composite variable, synthetic variable, derived variable, linear combination, super-variable, and weighted combination, to name a few. The variate represents the weighted combination of the values on the various predictor variables that will

correlate more highly with the criterion variable than any single predictor variable alone. Kachigan (1991) provided an appropriate education example, the final course grade. A typical course syllabus will explain that the final grade is composed of scores on different course requirements: quizzes, midterms, final exams, and/or theme papers. Because some requirements (variables) will contribute to the final grade more than others, the final grade is called a weighted combination.

In research, the advantage of these linear combinations of variables is that they increase the chances of discovering relationships (or differences) that single variable designs could not discern. If the multiple variables make theoretical sense, multivariate techniques are more powerful (Benton, 1991). Cooley and Lohnes (1976) assumed canonical correlation analysis “is the simplest model that can begin to do justice to this difficult problem of scientific generalization” (p. 176) with structural equation modeling subsuming canonical correlation analysis (Thompson, 2000). Although CCA and SEM belong to the same statistical phylum, they are distinct analyses with different goals. The primary goal of CCA is to maximize the relationship between the latent variables not to model the individual variables (Thompson, 1984). SEM, however, provides greater flexibility for the researcher to test the structure coefficients (Fan, 1997).

Purpose of Study

The purpose of this study was to compare the results of CCA to SEM on a heuristic data set. The correlation matrix along with the means and standard deviations are provided in Appendix A for interested readers to replicate the analyses. The participants for this study were 309 third-graders enrolled in the Alabama public school system. This data set is comprised of five independent variables (predictors) representing Piagetian tasks (classification, conservation, imagery, proportional reasoning, and relations). The five variables are measured by An Inventory of Piaget's Developmental Tasks (IPDT). The three dependent (criterion) variables are academic performance in language, reading, and math as assessed by the Alabama Basic Competency Tests (BCT). The hypothesis states there is a significant multivariate relationship between cognitive developmental levels as measured by the IPDT and competency as assessed by the BCT. This heuristic data set will be used to illustrate that structural equation modeling subsumes canonical correlation analysis. As Fan (1997) stated, "Hierarchically, the relationship between the two analytic approaches (canonical correlation analysis and structural equation modeling) suggests that SEM stands to be a more general analytic approach" (p. 65).

Statistical Analyses

At the most fundamental level, first and second-generation multivariate statistics share similar objectives. Kshirsagar (1972) described these objectives when he said “most of the practical problems arising in statistics can be translated, in some form or the other, as the problem of measurement of association between two vector variates \underline{x} and \underline{y} ” (p. 281). A brief description of the two statistical methods will be described subsequently.

Canonical Correlation Analysis

Canonical correlation analysis (CCA) is a member of the general linear model (GLM) family. CCA analyzes the relation between two sets of measured variables (Tabachnick & Fidell, 2001). One set may be viewed as the predictor variables while the other set may be considered the criterion or outcome variable. Similar to multiple regression, where the predictor variables are combined to form a single variate (a linear combination), canonical correlation is used to develop variates for both the predictors and the outcomes. Because the measured variables utilized for the predictor or criterion variables may represent different dimensions, more than one linear combination may emerge. The maximum number of variates will equal the number of variables in the smaller set (i.e., in this example, three pairs of variates will be formed).

Criticisms of CCA

Because CCA is considered an exploratory (first generation) statistical method (Tabachnick & Fidell, 2001), the shortcomings associated with exploratory methods need to be addressed. Nunnally (1978) noted that exploratory methods are neither "a royal road to truth, as some apparently feel, nor necessarily an adjunct to shotgun empiricism, as others claim" (p. 371). Exploratory results may not reproduce the relationships among the variables in another data set. The purpose of canonical correlation is to explain the relation of the two sets of variables, not to model the individual variables. This circumstance leads to the problem in the interpretation of the variates (Stevens, 2001). Mulaik, James, Van Alstine, Bennett, Lind, and Stilwell (1989) suggest that the difficulty in interpretation often comes about because the researcher lacks prior knowledge and therefore has no basis on which to make an interpretation.

Structural Equation Modeling

In contrast to CCA, Structural Equation Modeling (SEM) is a statistical technique that evaluates the plausibility of a hypothesized model. The full structural model can be decomposed into the structural model and the measurement model. The structural model assesses the relationships among the latent construct variables, which, in this example, are based on the Piagetian tasks (classification, conservation, imagery,

proportional reasoning, and relations) to academic performance (language, reading, and math). These latent constructs are usually defined by three to five measured variables. These multiple measures (the measurement model) allow the researcher to control more effectively for the inevitable measurement errors of any construct. By controlling for measurement error, unbiased estimates of the relationships among the latent constructs are possible. Once a model is proposed (i.e., relationships among the variables have been hypothesized), a correlation/covariance matrix is created. The estimates of the relationships among the variables in the model are calculated utilizing maximum likelihood estimation (MLE). MLE attempts to estimate the values of the parameters that would result in the highest likelihood of the actual data to the proposed model. These methods often require iterative solutions. With small samples, MLE may not be very accurate.

The model is then compared to the relationships (the correlation/covariance matrix) of the actual or observed data. SEM assesses how well the predicted interrelationships among the variables match the interrelationships among the actual or observed interrelationships. SEM assesses the measurement model (how well the measured variables define their respective construct) and the structural model (how well the latent constructs relate to each other) simultaneously. If the two matrices (the proposed and the actual or data) are consistent with one another, then the

structural equation model can be considered a credible explanation for the hypothesized relationships.

Stevens (2001) divided the model assessment into two categories: “those that measure the overall fit of the model and those that are concerned with individual model parameters” (p. 402-403). In SEM the overall fit of a model to the data may appear acceptable, yet some relations in the model may not be supported by the data. For example, an acceptable fit index could be achieved because of the strong measurement model though the structural model is fairly weak. Alternatively, the structural model may be impressive, but the measurement model may be quite weak, making the interpretation meaningless.

To assess the fit of the proposed model to the actual data, at least 24 fit indices have been proposed over the past 20 years (Klem, 2000). For the 24 SEM fit measures, there is presently no general agreement on which measures are preferred. As Hair, Anderson, Tatham, and Black (1998) stated, “SEM has no single statistical test that best describes the ‘strength’ of the model’s predictions” (p. 653). None of the measures has a related statistical test, except for the chi-square test (Hair et al, 1998). The fit indices are often classified into three types: (a) absolute fit measures (b) incremental fit measures, and (c) parsimonious fit measures. According to Maruyama (1997), “The different fit indexes differ with respect to dimensions such as susceptibility to sample size differences, variability in

the range of fit possible for any particular data set, and valuing simplicity of model specification needed to attain an improved fit" (p. 239). A brief description of common fit measures from the three categories is presented in the next section.

Absolute Fit Indices

The three most common absolute fit measures are the chi-square, the goodness-of-fit (GFI), and the root mean square error of approximation (RMSEA). The chi-square statistic is utilized to test the difference between the predicted and the observed relationships (correlations/covariances). Because the researcher is predicting a close fit, a non-significant chi-square is desired. The chi-square test, however, is often too powerful. As sample size increases, power increases. Therefore, the chi-square test can detect small discrepancies between the observed and predicted covariances and suggest that the model does not fit the data. A good fitting model could be rejected due to small differences between the observed and predicted value. Because of these limitations, other fit indices were developed as alternatives to the chi-square.

The goodness-of-fit index (GFI) is conceptually similar to the R^2 in multiple regression (Stevens, 2001). It is the proportion of variance in the sample correlation/covariance accounted for by the predicted model with values ranging from 0 (no fit) to 1 (a perfect fit).

The root mean square error of approximation (RMSEA) is the average of the residuals between the observed correlation/covariance from the sample and the expected model estimated from the population. Byrne (1998) stated, “(RMSEA) has only recently been recognized as one of the most informative criteria in covariance structure modeling” (p. 112). Values less than .08 are deemed acceptable while values greater than .10 are unacceptable.

One disadvantage of these absolute fit measures is that no distinction is made as to whether the model fit is better or worse in the structural or measurement models (Klem, 2000) (e.g., a model could demonstrate construct validity through the measurement model yet fail to demonstrate any statistically significant relationships among the latent variables in the structural model).

Incremental Fit Indices

Incremental fit measures are also known as comparisons to baseline measures. These are measures of fit relative to the independence model, which assumes that there are no relationships in the data. The independence model is the worst possible model. These measures, with values ranging from 0 to 1, indicate how much better the hypothesized model fits in comparison to the baseline that assumes that there are no relationships in the data. Values of .95 or greater are deemed acceptable.

Byrne (1994) suggested that the comparative fit index (CFI) should be the fit statistic of choice in SEM analyses.

Parsimonious Fit Indices

In general, researchers are seeking not only the best fitting but also the most parsimonious model. Two such fit indices are the parsimony adjusted CFI (PCFI) and the parsimony adjusted GFI (PGFI). These fit indices adjust the estimate to provide a comparison between models with different numbers of estimated parameters to determine the impact of adding additional parameters to the model. No statistical test, however, is available for these fit indices, which limits their use to model comparisons.

Comparing SEM to CCA

SEM has advantages over first-generation techniques such as canonical correlation because of the greater flexibility for the researcher to compare observed data with a theory. SEM involves generalizations and extensions of first-generation procedures (Chin et al., 1996). In fact, SEM analysis would yield a first-generation analysis if certain constraints were applied. Advantages to using SEM to compute canonical results, statistical significance testing of individual canonical function coefficients and structure coefficients is possible, and individual canonical coefficients can

be tested (Fan, 1997). This assessment is not easily accomplished in conventional canonical analysis (Thompson, 1984).

Results

Canonical Correlation

Canonical correlation was used to examine the multivariate associations between the Piagetian tasks as measured by the IPDT and competency in reading, language, and mathematics as measured by the BCT. Canonical correlation revealed one significant dimension relating the two sets of variables: $r = .53$, $F(15, 831) = 7.81$, $p < .01$. Correlations of each of the measures with the canonical variate (structure coefficients) are provided in Table 1. Correlations of .30 and higher are statistically significant (Tabachnick & Fidell, 2001). These results suggest that there is a common dimension underlying cognitive development and school performance.

Structural Equation Modeling

Using AMOS 4.0 (Arbuckle, 1999), the relationship between the IPDT, a latent variable with five indicators (classification, conservation, imagery, proportional reasoning, and relations), and the BCT, a latent variable with three indicators (reading, language, and math), was examined. Although the chi-square test was significant, $\chi^2(19) = 50.82$, $p < .01$, the model yielded acceptably high goodness of fit indices (.998 and .994) for both the CFI and

Table 1.
Structure Coefficients for Canonical Correlation and Structural Equation Modeling

Variable	Canonical Correlation	Structural Equation Modeling
(IPDT)		
Classification	0.73	0.70
Conservation	0.62	0.61
Imagery	0.64	0.58
Proportional Reasoning	0.65	0.65
Relations	0.89	0.72*
(BCT)		
Reading Score	0.85	0.94
Language Score	0.61	0.91*
Math Score	0.88	0.89

*Statistically significantly different at $p < .05$.

the TLI respectively. Joreskog and Sorbom (1989) and Bentler (1990) warned against the sole use of the chi-square value in assessing the fit of the model because of the sensitivity of the chi-square to sample size. The RMSEA achieved a value of .074 indicating an acceptable fit of the model in relation to the degrees of freedom. All measured variables loaded on their respective factors significantly ($p < .01$). The structural coefficients are

Guarino

presented in Table 1. The IPDT correlated to the BCT significantly ($r = .51, p < .01$).

The structural path and loadings (structure coefficients) were of considerable strength as opposed to just being statistically significant (Table 1). All of the structure coefficients approached 0.60 indicating that each measure is accounting for a substantial percentage of the variance of the underlying latent variable. The standardized path coefficients should be above .30 to be considered meaningful. Meehl (1990) argued that anything lower might be due to what he has termed the crud factor where “everything correlates to some extent with everything else” (p. 204).

Comparing Coefficients

To test if the structure coefficients are invariant between the two analyses (CCA and SEM), Hotelling’s t test was employed because the correlations were obtained from the same sample. No significant differences were reported for the relationship between IPDT and BCT (.51 and .53) for the SEM and CCA respectively. Significant differences were reported on the loadings for Relations on the IPDT (.72 for SEM, .89 for the CCA) and Language on the BCT (.91 for SEM, .61 for the CCA).

Discussion

Concurrent validity between the two latent variables (IPDT and BCT) was supported by both the exploratory and confirmatory analyses. The CCA derived one significant dimension relating the two sets of variables, which was hypothesized in the SEM procedure. The canonical root was 28%, which was slightly greater than the variance explained from the SEM procedure, which was 26%. This difference in variance explained may be that CCA derived the maximum relationship between the IPDT and the BCT, in this case a nonsignificant 2% increase.

Although the relationship between IPDT and BCT was invariant for the two methods, the corresponding factor structures for the two latent variables were variant. In this example, the CCA overestimated (relative to SEM) the contribution of Relations on the IPDT variate and underestimated the contribution of Language on the BCT variate. Thus, the factor structures were variant as derived by the different statistical analyses, which leads to the problem in the interpretation of the respective variates. These differing factor structures are consistent with the goals of the two methods (SEM and CCA) examined in this study. Because the primary goal of CCA is to maximize the relationship between the latent variables, CCA will develop structure coefficients for their respective variates accordingly.

In summary, the advantages of SEM over CCA were illustrated in this study with SEM providing a more accurate representation of the latent

construct structure. Because the decision as to which statistical test to use (CCA or SEM) is contingent in part on the current understanding of the phenomenon being investigated, SEM was the more appropriate choice. Both the predictors (the Piagetian tasks measured by The Inventory of Piaget's Developmental Tasks) and the outcome (academic performance in language, reading, and mathematics assessed by the Alabama Basic Competency Tests) were grounded in a theoretical framework. Recall that the purpose of canonical correlation is to explain the relation of the two sets of variables, not to model the individual variables. The assessment of the structure coefficients is not easily accomplished in CCA (Thompson, 1984). This study, through the use on a heuristic data set, illustrated the advantages of SEM over CCA because SEM provided the researcher a more accurate assessment of the structure coefficients.

References

- Arbuckle, J. L. (1999). *Amos 4.0 user's guide*. Chicago: SmallWaters Corporation.
- Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, 107, 238-246.
- Benton, R. (1991). Statistical power considerations in ANOVA. In B. Thompson (Ed.), *Advances in educational research: Substantive findings, methodological developments* (Vol. 1., pp. 119-132). Greenwich, CT: JAI Press.
- Byrne, B. M. (1994). *Structural equation modeling with EQS and EQS/Windows: Basic concepts, applications, and programming*. Thousand Oaks, CA: Sage Publications.

- Byrne, B. M. (1998). *Structural equation modeling with LISREL, PRELIS, and SIMPLIS*. Hillsdale, NJ: Lawrence Erlbaum.
- Chin, W. W., Marcolin, B. L., & Newsted, P. R. (1996). A partial least squares latent variable modeling approach for measuring interaction effects: Results from a Monte Carlo simulation study and voice mail emotion/adoption study. In J. I. DeGross, S. Jarvenpaa, & A. Srinivasan (Eds.), *Proceedings of the Seventeenth International Conference on Information Systems* (pp. 21-41).
- Cooley, W. W., & Lohnes, P. R. (1976). *Evaluation research in education*. New York: Irvington.
- Fan, X. (1997). Canonical correlation analysis and structural equation modeling: What do they have in common? *Structural Equation Modeling*, 4(1), 65-79.
- Fornell, C. (1987). A second generation of multivariate analysis: Classification of methods and implications for marketing research. In M. J. Houston (Ed.), *Review of Marketing*, (pp. 407-450). American Marketing Association, Chicago, IL.
- Hair, J.F., Anderson, R.E., Tatham, R.L., & Black, W.C. (1998). *Multivariate Data Analysis 5th Ed.* Upper Saddle River, NJ: Prentice-Hall, Inc.
- Kachigan, S.K. (1991). *Multivariate statistical analysis: A conceptual introduction* (2nd ed.). New York: Radius Press.
- Klem, L. (2000). Structural equation modeling. In L.G. Grimm & P.R. Yarnold (Eds.), *Reading and understanding more multivariate statistics* (pp. 227-260). Washington, DC: American Psychological Association.
- Kshirsagar, A.M. (1972). *Multivariate analysis*. New York: Marcel Dekker.
- Joreskog, K.G., & Sorbom, D. (1989). *LISREL 7: A guide to the program and applications* (2nd ed.). Chicago: SPSS.
- Maruyama, G. M. (1997). *Basics of structural equation modeling*. Thousand Oaks, CA: Sage Publications.
- Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterruptible. *Psychological Reports*, 66, 195-244.

- Mulaik, S. A., James, L. R., Van Alstine, J., Bennett, N., Lind, S. & Stilwell, C. D. (1989). Evaluation of goodness-of-fit indices for structural equation models. *Psychological Bulletin*, 105, 430-445.
- Nunnally, J. (1978). *Psychometric theory* (2nd ed.). New York: McGraw Hill.
- Stevens, J. (2001). *Applied multivariate statistics for the social sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.
- Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics* (4th ed.). Needham Heights, MA: Allyn & Bacon.
- Thompson, B. (1984). *Canonical correlation analysis: Uses and interpretation*. Newbury Park, CA: Sage.
- Thompson, B. (1991). A primer on the logic and use of canonical correlation analysis. *Measurement and Evaluation in Counseling and Development*, 24, 80-95.
- Thompson, B. (2000). Ten commandments of structural equation modeling. In L.G. Grimm & P.R. Yarnold (Eds.), *Reading and understanding more multivariate statistics* (pp. 261-283). Washington, DC: American Psychological Association.

Appendix

Correlation matrix of the IPDT and BCT

	CLASS	CONS	IMAG	PROR	RELS	READTOT	LANGTOT
CONS	.404						
IMAG	.385	.410					
PROR	.482	.412	.339				
RELS	.501	.424	.413	.466			
READTOT	.346	.263	.320	.316	.408		
LANGTOT	.270	.181	.257	.234	.266	.866	
MATHTOT	.372	.301	.331	.315	.407	.826	.815

Note. $N = 309$.

Univariate Descriptive Statistics

Variable	M	S
CLASS	7.36	2.68
CONS	6.76	2.54
IMAG	6.80	2.27
PROR	6.47	2.65
RELS	6.06	2.34
READTOT	77.16	11.74
LANGTOT	61.96	8.76
MATHTOT	60.13	7.36

Note. $N = 309$.