DOCUMENT RESUME

ED 302 571 TM 012 576

AUTHOR Baldwin, Janet
TITLE Validating the Factor Structure of Ratings Assigned to Essays: A Confirmatory Factor Analytic Approach.
PUB DATE Apr 88
NOTE 20p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 5-9, 1988).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Essay Tests; *Factor Analysis; Goodness of Fit; Grade 9; Grade 10; High Schools; Mathematical Models; State Programs; *Test Validity; Writing (Composition); *Writing Evaluation; Writing Skills
IDENTIFIERS *Unidimensionality (Tests)

ABSTRACT

The use of confirmatory factor analytic procedures to examine the dimensionality of writing skills as measured by a large-scale direct writing test was illustrated. Internal construct validity evidence about the nature of writing skills measured by the test was provided. Data used were scores assigned by about 100 trained professional raters on a state-wide competency-based test of writing skills. A total of 3,430 students responded to all four prompts of the writing tests. A sequential hypothesis testing strategy was followed for a one-factor model, a two-factor mode-specific model, and a two-factor occasion-specific model. These were compared to a null model specifying no common factors. Results from these analyses provided evidence about the usefulness of the construct-based hypotheses for explaining the variability in data from this test. A one-factor model provided a more plausible and parsimonious representation of the data than did the two-factor hypotheses. Evidence in support of a unidimensional interpretation of the scores suggests there may be little practical distinction between narrative and explanatory scores. A table displays goodness-of-fit statistics. (SLD)

Validating the Factor Structure of Ratings Assigned to Essays: A Confirmatory Factor Analytic Approach

Janet Baldwin
American Council on Education

Presented at the Annual Meeting of the American Educational Research Association, New Orleans, April 1988

Validating the Factor Structure of Ratings Assigned to Essays: A Confirmatory Factor Analytic Approach

Janet Baldwin

American Council on Education

An important prerequisite to the study of relationships between constructs is the study of the nature of the constructs themselves and their effects on observed measures (Costner, 1969; Campbell & Fiske, 1959). Studying the linkage between constructs and their indicators addresses questions about construct validity of measurements. Such studies provide valuable information for researchers and psychometricians seeking to compare the meaning of a test's scores across persons, occasions, items and domains. For example, as Floden, Porter, Schmidt, and Freeman (1980) pointed out, a test may be viewed as unidimensional based on internal consiscency evidence. However, it may not be unidimensional for all types of respondants, under various conditions, at different times, or over different settings (Floden et al., 1980).

Confirmatory factor analysis (CFA) is a powerful method for examining the measurement aspects of structural models (Long, 1983); Jöreskog, 1979b). As Dwyer (1983) has pointed out, the real power of these procedures for testing measurement models comes from the ability to specify empirical predictions that flow from competing models which may explain the variability among a set of indicators. Estimation techniques then can be used to evaluate whether one model is superior to another.

Theoretical Framework

Various theories about writing skill address such important questions as whether writing skill varies according to the topic, the requested mode, or the particular writing occasion. For example, in at least one construct validation study (Breland, Camp, Jones, Morris, and Rock, 1987), writing ability based on essay data was viewed as a single generalizable construct as well as in terms of multiple factors which may be rater-, content-, mode-, or topic-specific. If writing samples in narrative and in explanatory modes each elicit conceptually distinct types of writing, a model representing two-mode factors should fit better than a model representing a single general writing factor, thereby supporting the discriminant validity of separate mode scores. Finding a better fit for a model specifying a separate factor for each test occasion, however, would raise questions about either the generality, or pervasiveness, of the factors measured by the testing procedure, the comparability of the test forms on those occasions, or both, creating potential problems in the interpretation of test scores. Such internal construct validation studies not only provide important insights into the nature of what is measured, but they are important prerequisites to subsequent external validations studies.

This study had two objectives. The first objective was to illustrate the use of confirmatory factor analytic procedures to examine the dimensionality of writing skills as measured by a large-scale direct writing test. The second objective was to provide internal construct validity evidence regarding the nature of the writing skills measured by such a test. Specifically, the

study sought to determine whether a one-factor model or a two-factor model provided a better fit to the essay rating data. Three types of models were specified, of which two were derived from theories about writing constructs. That is, a one-factor theoretical model was specified to represent general writing ability and a two-factor theoretical model was specified to represent two related but conceptually distinct domains of writing skill, narrative and explanatory writing. To better evaluate the results from these two types of models a third type of model, derived from an alternative, or competing, view, was tested. This model specified a separate factor for each of two test administration occasions. Models with more than two-factors may be specified to represent scores assigned to essays, such as a factor for each topic. However, because only two ratings were available for each topic in this study, a model with separate topic-specific factors may not have been identified. Therefore, for the purposes of this study, the models tested were limited to models with only one or two-factors.

Instrument and Data Source

Data used for this study were holistic scores assigned by trained professional raters to writing samples of high school students on a state-wide competency based test in writing skills. The writing test used in this program consisted of two prompts, one narrative and one explanatory, which were administered to ninth and tenth graders throughout the state as part of the state's requirements for a high school diploma. Each year, different test forms are administered, along with an anchor test

form which is administered for equating purposes.

In this study, data from two writing test forms were used. Form A, the anchor test form, consisted of two prompts, one designed to elicit narrative writing (N_1) and another designed to elicit explanatory writing (E_1). Form B, which was designed to be parallel to Form A, consisted of another pair of prompts, one narrative (N_2) and one explanatory (E_2). The prompts used in each operational test were selected for their content and statistical comparability to previous forms. That is, test forms were constructed from prompts which elicited score distributions under field test conditions that were similar to those of previous tests. Therefore, these forms may be viewed as essentially parallel.

Form B (N_2 and E_2) was administered to high school students in grades 9 and 10 during a ten day period in January 1987. In addition, 3,430 students from ten schools in six school districts also composed responses to the prompts from Form A (N_1 and E_1) of this writing test as part of an annual equating study. The equating sample was divided into two groups so that the two prompts in Form A were administered to one group before the administration of Form B and to the other group after the administration of Form B. Scores were assigned by approximately 100 professional readers in each domain who were trained in modified holistic procedures to apply either a four point narrative scale or a four point explanatory scale to students' narrative or explanatory writing samples. Within each domain, each writing sample was scored independently by two raters, providing two scores for each of four prompts. Only those raters

of demonstrated scoring ability were perimitted to assign scores to the essays. A covariance matrix of these eight scores was used as input data for the LISREL CFA analyses described below.

Of the 3,430 students who responded to all four prompts, 50% were female and 50% were male. Of the 73% who indicated their race or ethnic group, 59% were White, 27% were Black, 11% were Hispanic, and 3% were Asian. Complete data on all four prompts (eight essay scores) were available for 3002 students.

For Form B, score reliability coefficients (coefficient alpha) were .94 and .96, respectively, for the narrative and the explanatory topics. Exact agreement rates (percents) on the narrative prompts and the explanatory prompts were 76% and 72%, respectively. When combined with adjacent agreement rates, the total rater agreement rates were over 99% (Maryland State Department of Education, 1987). For Form A, these statistics were unavailable. However, data from field testing indicated that within each domain, the score distributions for these two forms were highly similar and judged to be essentially parallel. Therefore, it was assumed that both forms have highly similar psychometric properties within the narrative and explanatory domains.

Methods

Exploratory factor analytic procedures often are applied to data as a prior stage to confirmatory factor analysis because CFA requires substantive knowledge about the data in order to specify plausible and testable hypotheses. In the case of these data, it was known what domains of writing the data were intended to

measure as well as the rules by which scores were assigned. Therefore, it was possible to specify substantively meaningful hypotheses for empirical tests without examining the structure of the data through a preliminary exploratory factor analysis.

Following Jöreskog (Jöreskog & Sörbom, 1983), it is assumed that a factor analysis model holds for the essay rating data, so that

$$\underset{\sim}{x} = \underset{\sim}{\Lambda} \cdot \underset{\sim}{\xi} + \underset{\sim}{\delta} ,$$

with parameter matrices for factor loadings (Lambda), latent factor variances and covariances (Phi), and measurement errors (Theta Delta).

A sequential hypothesis testing strategy was followed for each of the following three models: a one-factor model, a two-factor mode-specific model, and a two-factor occasion-specific model. For the one-factor model, all essay ratings obtained on two topics (narrative and explanatory) over two-occasions were specified to load on a single writing skill factor. For the mode-specific model, essay ratings of narrative writing were specified to load on a narrative writing factor and essay ratings of explanatory writing were specified to load on an explanatory writing factor. For the occasion-specific model, essay ratings of both modes of writing assigned in one test administration were specified to load on an Occasion 1 factor, and essay ratings obtained in the other test administration were specified to load on an Occasion 2 factor.

These one-and two-factor models were compared to a null model which specified no common factors. This provided a means for computing some of the indices of fit by which relative

improvements were evaluated. In addition, other comparisons of model-data fit were made to determine the best baseline model for each theoretical view. Within each of the three models, hypotheses were tested to determine whether the measurement errors were correlated for essay ratings of the same writing topic. This comparison sought to determine whether a model which specified correlated measurement errors between the ratings for each topic provided a significant improvement in fit over a model specifying uncorrelated measurement errors between ratings. Because it is reasonable to expect that measurement errors associated with two ratings of the same writing sample may be correlated, it was important to test this expectation in order to provide a plausible baseline model against which to compare subsequent restricted models.

In these analyses, all latent factors were constrained to have unit variances, error variances were freely estimated, and for the two-factor models, correlations between latent factors were estimated. In addition, the following sequence of increasingly restricted hypotheses was tested for equality of factor loadings:

Hypothesis A: across ratings within topic

Hypothesis B: across all ratings within each mode

Hypothesis C: across all ratings within each occasion

CFA procedures (Jöreskog & Sörbom, 1983) were applied to specify and test these models. Criteria for model data fit included chi-square difference tests (Jöreskog & Sörbom, 1983) and several indices of fit produced as output in the LISREL (Jöreskog

& Sörbom, 1983) program: chi-square (probability) value, Goodness of Fit index (GFI), Root Mean Square Residual (RMR), Modification Index (MI), and Normalized Residuals (NR). In addition, the Parsimonious Fit Index (PFI) (James, Mulaik, and Brett, 1982) and ratios of chi-square to degrees of freedom (Jöreskog & Sörbom, 1983) were used. Finally, judgements were made about which of the models provided the most plausible and parsimonious representation of the data.

Descriptions of the goodness of fit indices produced in the LISREL program are found in Jöreskog and Sörbom (1983). The parsimonious fit index (PFI) is actually Bentler and Bonett's (1980) normed fit index modified to take into account the number of degrees of freedom given up in order to arrive at a particular level of goodness of fit. Generally, the models with the maximum values of PFI are those that best describe the data with the fewest unknown parameters (Loehlin, 1987). The range for recommended ratios of chi-square/degrees of freedom (df) typically are between 2 and 5 (Carmines & McIver, 1981).

Because the chi-square value is dependent on sample size, the chi-square probability value is likely to be significant in large samples regardless of how well the model fits the data, whereas in small samples it may be non-significant even for models which are poor. Therefore, the chi-square probability value can be misleading, thereby reducing its usefulness as an indicator of fit. Far more relevant as criteria of fit are sequential tests of incremental differences in fit, or chi-square difference tests, because such tests improve inference with both large and small samples (Bentler, 1980). Because the differences in chi-square

values are themselves chi-square statistics, they can be used to test the importance of parameters that differentiate nested models.

Results and Conclusions

The chi-square probabilities for all analyses were highly significant. Other goodness of fit results from each analysis are presented in Table 1. Generally, the null model (Model 1) represents the most restricted model against which other less restricted models may be compared. In the analyses reported below, the null model provided a basis for calculating the PFI. Within each theoretical view, baseline models were selected and chi-square difference tests were used to evaluate the influence of subsequent restrictions on model data fit.

As shown in Table 1, the one- and two-factor models (Models 2, 7, and 11) provided a considerable improvement in fit over the null model (Model 1). In addition, baseline comparisons were made to determine the influence of correlated measurement errors between within-topic ratings of essays on the fit of the one-factor and two-factor models.

The one- and two-factor models without correlated errors (Models 2, 7, and 11) then were compared to one- and two-factor models with correlated errors (Models 3, 8, and 12). Within each of the theoretical views represented by these three model types, the correlated error models provided significant improvements in fit. For the one-factor model, chi-square (df) dropped from 3,188.48 (20) (Model 2) to 55.13 (16) (Model 3); for the two-mode factor model, chi-square (df) dropped from 2,475.72 (19) (Model 7)

to 41.89 (15) (Model 8); for the two-occasion factor model, chi-square (df) dropped from 2,519.79 (19) (Model 11) to 46.52 (15) (Model 12).

Not only are these chi-square differences highly significant, but other fit indices also indicated good fit. For Models 3, 8, and 12, the GFI values were all above .996 and the RMR values were all less than .009. An examination of the normalized residual matrix for these models indicated that none were greater than 2.0. The chi-square/degrees of freedom ratios were 3.44, 2.79, and 3.10 for Models 3, 8, and 12, respectively. Generally, ratios between 2 and 5 are considered acceptable, with those closer to 2 indicative of a better fit. Therefore, while all were within acceptable ranges, the two-factor models provided slightly better fits than the one-factor model, and the two-mode factor model (Model 8) was slightly better than the two-occasion factor model (Model 12). Modification indices (MI) were also examined for these models. The MI indicates the amount of decrease in chi-square which would result if the parameter were freed. When the maximum values of the MI are relatively small, it suggests that the model can be improved very little by freeing up additional parameters. The maximum modification indices were 4.28, 3.46, and 3.70 for Models 3, 8, and 12, respectively. The relatively small magnitudes of these MI values suggest that the models will be improved very little by freeing up additional parameters. Because all these indices of fit are quite good, it can be concluded that the better baseline models for the three theoretical views are those which contain correlated errors between ratings of essays on the same topic. Chi-square difference tests then were applied to

compare the influence of additional restrictions placed on the model.

One Factor Models

When factor loadings for the one-factor model (Model 3) were constrained to be equal within each topic (Model 4), the chi-square (df) increased from 55.13 (16) to 64.32 (20). This chi-square difference of 9.19 (4) is not signficant. Therefore, by constraining the factor loadings within each topic to be equal, the one-factor model (Model 4) produced a fit not significantly different from the fit obtained without these constraints. Moreover, the fit for Model 4 was more parsimonious than Model 3. Chi-square difference tests were also applied to compare the fit of Model 3 with models constraining the factor loadings to equality within each mode (Model 5) and within each occasion (Model 6). In both comparisons, the differences -- 18.46 (4) and 55.40 (6), respectively -- were significant, which means that the model-data fit was significantly poorer when these equality constraints were applied. Therefore, the best fitting one-factor model was Model 4, which had correlated errors between ratings of essays on the same topic and factor loadings within each topic constrained to be equal.

Two Mode Factor Models

When factor loadings for the two-mode factor model (Model 8) were constrained to be equal within each topic (Model 9), the chi-square (df) increased from 41.89 (15) to 50.78 (19). The chi-square difference of 8.89 (4) is not significant, suggesting

that the additional equality constraints on the factor loadings within each topic resulted in a fit not significantly different from the fit obtained without these constraints. Because fewer parameters are estimated in this model, Model 9 represents a more parsimonious representation of the data than does Model 8. When factor loadings for the two-mode model (Model 8) were constrained to be equal within each writing mode (Model 10), the chi-square (df) increased from 41.89 (15) to 60.44 (21). This is a significant difference in chi-square and represents a significantly poorer fit for Model 10. Therefore, the best fitting two-mode factor model was Model 9, which had correlated errors between ratings of essays on the same topic and equality constraints on factor loadings within each writing topic.

When factor loadings for the two-occasion factor model (Model 12) were constrained to be equal within each topic (Model 13), the chi-square (df) increased from 46.52 (15) to 55.90 (19). The chi-square difference of 9.38 (4) is not significant. Therefore, the additional equality constraints on the within topic factor loadings result in a model-data fit which is not significantly different from the fit obtained without these constraints. As for the two-mode factor model, fewer parameters are estimated in Model 13, representing a more parsimonious representation of the data. When the factor loadings for the two-occasion model (Model 12) were constrained to be equal within each writing occasion (Model 14), the chi-square increased from 46.52 (15) to 101.19 (21), a highly significant difference. Therefore, the best fitting two occasion factor model was Model 13, which had correlated errors

between ratings of essays on the same topic and equality constraints on factor loadings within each topic.

The other fit indices for Models 4, 9, and 13 indicate satisfactory fits relative to the other models. For example, the chi-square/df ratios for Models 4, 9, and 13 provide the lowest ratios within each model type, with ratios of 3.22, 2.67, and 2.94, respectively. The GFI and RMR values are all within satisfactory ranges: for these three models the GFI values are .995 or more and the RMR values are .011 or less. None of the normalized residual matrices for these models have elements with values greater than 2.0. The PFI values for Models 4, 9, and 13 (.95, .95, and 1.00, respectively) indicate that each model provides a parsimonious fit to the data.

Best fitting model

Because both the one-factor and the two-factor models may be viewed as nested models (Loehlin, 1987), the relative improvement in fit of the two-factor models over the one-factor model can be examined through chi-square difference tests. The difference in chi-square values for Model 4 and Model 9 (64.32 - 50.78) is 13.54. With one degree of freedom, this is a significant difference. Therefore, the two-mode model (Model 9) represents a significant improvement over the one-factor model (Model 4). The difference in chi-square values for Model 4 and Model 13 (64.32 - 55.90) is 8.42. With one degree of freedom, this difference is also significant. Therefore, even the two-occasion model represents a statistically significant improvement over the one-factor model.

Because Models 9 and 13 both have the same degrees of freedom, a chi-square difference test cannot be applied. However, an absolute difference in chi-square of 5.12 is large enough to suggest the superiority of the two-mode model, Model 9. In addition, the relative differences between these two-models can be evaluated on the basis of the information provided by the other goodness of fit indices. The chi-square/df ratio for Model 9 (2.67) is smaller than those for Models 4 (3.22) and 13 (2.94). The GFI and RMR for these models are all nearly the same.

Based on statistical tests alone, the two-mode factor model (Model 9) would appear to provide the best fit to the data of the three best-fitting models within each model type. However, an inspection of the estimated correlation between the narrative and explanatory latent factors for Model 9 reveals a near unity correlation of .96. A near unity estimated correlation suggests that these two-factors may, in fact, be measuring the same general dimension of writing ability. Jöreskog (1979a) has emphasized the importance of substantive interpretability in evaluating statistical goodness of fit, pointing out that it is unwise to rely on statistical criteria alone. If the estimated correlation between the two writing modes is near unity, it seems unreasonable to conclude that the data reflect two conceptually distinct modes of writing. An examination of the parsimonious fit index for Model 4 (1.00) suggests that Model 4, the one-factor model with correlated errors between ratings of each topic, may provide a slightly more parsimonious fit than that provided by Model 9 (.95) and Model 13 (.95). Although the chi-square difference tests lead to the conclusion that a two-factor model is superior, a more

plausible and parsimonious representation appears to be provided by the one-factor model. Indeed nearly as good a fit can be obtained by one-factor model with correlated errors and equal factor loadings within topic as can be obtained by specifying two-factors with similar constraints.

Conclusions

Testing the fit of one-factor and two-factor models based on a general writing factor, on mode-specific factors, or on occasion-specific factors provided comparisons between a set of construct-based models and a set of alternative, or non-construct-based models. The results from these analyses provided evidence about the usefulness of construct-based hypotheses for explaining the variability in data from this test.

Because a one-factor hypothesis provided a more plausible and parsimonious representation of the data than did a two-factor hypothesis, there appears to be little support for the discriminant validity of separate writing mode scores. Indeed, the evidence in support of a unidimensional interpretation of the scores assigned by raters suggests that there may be little practical distinction between the narrative scores and the explanatory scores.

It is important to note that not all relevant factor models were tested with these data. For example, it may be reasonable to suggest that the data are represented by two primary factors related to narrative and explanatory writing, and a higher order factor related to general writing ability. As this model was not tested, it is recommended that such a model be examined in future

studies. It is important to point out that the distributional attributes of essay score data may violate to some extent the multivariate normality assumption required in the application of maximum likelihood procedures in LISREL. Therefore, results from these analyses should be interpreted with caution. Nevertheless, all LISREL analyses converged quickly, all standard errors were positive, t-values for all estimated parameters were very large, and there were no instances of improper solutions.

Confirmatory factor analytic procedures provide powerful tools for examining the relationship between important psychological constructs and their observed measures under various conditions. Because essay rating data could have many sources of variation which may be unrelated to the construct of interest, it is important to study the nature of the constructs which are measured by such procedures. Holistic scoring of essays is becoming widely used as a measure of writing skill and such measures often are used by teachers to provide feedback to students. Therefore, a better understanding of what is measured in essay ratings can lead not only to improved writing assessment but also to improved writing instruction. An empirical study of writing skill and its relationship to relevant indicators offers useful information about both the construct and the test variables. Finally, the study demonstrates the application of a useful methodology for examining the construct validity of an increasingly common approach to the measurement of writing skill.

Table 1. Goodness of Fit Statistics For Confirmatory Factor Analysis Models Of Ratings Assigned to Essays

Models	Chi-square	df	χ^2/df	GFI	RMR	NR>2.0	Max MI	PFI
1.Null	15,352.92	20	767.65	.313	.302			
2.One Factor	3,188.48	20	159.42	.794	.047	12	0	.79
3.1FCE	55.13	16	3.44	.996	.009	0	4.28	.80
4.1FCE-LT	64.32	20	3.22	.995	.011	0	4.79	1.00
5.1FCE-LM	73.59	22	3.34	.994	.015	4	9.13	1.09
6.1FCE-LO	110.53	22	5.02	.991	.028	11	11.72	1.09
7.Two Modes	2,475.72	19	130.30	.811	.039	12	2.81	.80
8.2MCE	41.89	15	2.79	.997	.008	0	3.46	.75
9.2MCE-LT	50.78	19	2.67	.996	.010	0	4.63	.95
10.2MCE-LM	60.44	21	2.88	.995	.014	3	9.46	1.04
11.Two Occasions	2,519.79	19	132.62	.808	.039	12	2.28	.79
12.2OCE	46.52	15	3.10	.996	.008	0	3.70	.75
13.2OCE-LT	55.90	19	2.94	.995	.011	0	4.57	.95
14.2OCE-LO	101.19	21	4.82	.991	.027	9	11.97	1.04

1. Null: No common factors, eight variables, perfectly measuring eight factors.

2. One Factor: one common factor, eight variables.
3. 1FCE: same as Model 1 with correlated errors between ratings on same prompt.
4. 1FCE-LT: same as Model 3 with lambdas within topics constrained to be equal.
5. 1FCE-LT: same as Model 3 with lambdas within modes constrained to be equal.
6. 1FCE-LO: same as Model 3 with lambdas within occasion constrained to be equal.

7. 2 Modes: two common factors, four narrative and four explanatory variables.
8. 2FCE: same as Model 7 with correlated errors between ratings on same prompt.
9. 2FCE-LT: same as Model 8 with lambdas within topics constrained to be equal.
10. 2FCE-LM: same as Model 8 with lambdas within modes constrained to be equal.

11. 2 Occasions: two common factors, four measures each for occasions 1 and 2.
12. 2FCE: same as Model 11 with correlated errors between ratings on same prompt.
13. 2FCE-LT: same as Model 12 with lambdas within topics constrained to be equal.
14. 2FCE-LO: same as Model 12 with lambdas within occasions constrained to be equal.

References

Bentler, P.M. (1980). Multivariate analysis with latent variables: Causal modeling. In M.R. Rosenzweig and L.W. Porter (Eds.), Annual Review of Psychology, 31, 419-456.

Bentler, P.M. and Bonett, D.G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. Psychological Bulletin, 88, 588-606.

Breland, H., Camp, R., Jones, R., Morris, M. and Rock, D. (1987). Assessing Writing Skill. Research Monograph No. 11. College Entrance Examination Board, New York.

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

Carmines, E. and McIver, J. (1981). Analyzing models with unobserved variables: Analysis of covariance structures. In G. Bohrenstedt and E. Borgatta (Eds.), Social measurement: Current issues. Beverly Hills:Sage.

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

Dwyer, J. Statistical models for the social and behavioral sciences. New York: Oxford University Press, 1983.

Floden, R., Porter, A., Schmidt, W., and Freeman, D. (1980). Don't they all measure the same thing? Consequences of standardized test selection. In E.L. Baker and E.S. Quellmalz (Eds.), Educational testing and evaluation: Design, analysis, and policy. Beverly Hills:Sage.

James, L.R., Mulaik, S.A., & Brett, J.M. (1982). Causal analysis: Assumptions, models, and data. Beverly Hills:Sage.

Jöreskog, K. (1979a). A general approach to confirmatory maximum likelihood factor analysis. In K.G. Jöreskog and D. Sörbom (Eds.), Advances in factor analysis and structural equation models. Cambridge, Mass.:Abt.

Jöreskog, K. (1979b). Analyzing psychological data by structural analysis of covariance matrices. In K.G. Jöreskog and D. Sörbom (Eds.) Advances in factor analysis and structural equation models. Cambridge, Mass.:Abt.

Jöreskog, K. and Sörbom, D. (1983). Analysis of linear structural relations by the method of maximum likelihood. Users Guide, Version VI (2nd Ed.). Chicago:International Educaitonal Services.

Loehlin, J.C. (1987). Latent variable models: An introduction to factor, path, and structural analysis. Hillsdale, N.J.:Lawrence Erlbaum Associates.

Long, J. (1983). Confirmatory factor analysis. Quantitative applications in the social sciences. Beverly Hills:Sage.

Maryland State Department of Education (1987). Technical Report: Maryland Functional Writing Test-II, Winter Administration. Baltimore, Md.:Author.