DOCUMENT RESUME

ED 322 168 TM 015 287

AUTHOR	Dabney, Marian E.; Stewart, Theadora
TITLE	A Validation Study of a Georgia Teacher Certification Test Using Confirmatory Factor Analysis.
PUB DATE	Apr 90
NOTE	32p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Boston, MA, April 17-19, 1990).
PUB TYPE	Reports - Research/Technical (143) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Computer Assisted Testing; Computer Software; *Construct Validity; Content Validity; Factor Analysis; *Licensing Examinations (Professions); Maximum Likelihood Statistics; Psychometrics; Sample Size; *Special Education Teachers; *State Programs; *Teacher Certification; Testing Programs; *Test Validity
IDENTIFIERS	Confirmatory Factor Analysis; Georgia; LISREL Computer Program; One Parameter Model

ABSTRACT

This study investigated the construct validity of the revised Special Education-Mental Handicaps Georgia Teacher Certification Test (MH-TCT) using hierarchical confirmatory factor analysis and LISREL VI. The primary objective was to determine whether first-order and second-order factors correspond to item/objective/test relationships defined by judgments/consensual content validation studies. Data from four administrations of the MH-TCT were used in the analyses. Subjects included 287 candidates for certification in the mental handicaps subarea of special education. Since the sample size was small, findings were tentative. LISREL was used to examine the factor structure of the entire test as well as to identify test items that were not psychometrically sound. Although six subdimensions were identified that correspond to the six test objectives, first-order factors corresponding to these were so highly correlated that a hierarchical model was rejected in favor of a unidimensional model. A unidimensional model was invariant across samples and time. Results suggest that the LISREL confirmatory maximum likelihood factor model provided a powerful tool for item analysis and test validation. A major drawback was the amount of computer run time needed to do maximum likelihood factor analysis. Four data tables and one flowchart are provided. (Author/TJH)

A Validation Study of a Georgia Teacher Certification Test Using Confirmatory Factor Analysis

Marian E. Dabney

Georgia State Department of Education

Theadora Stewart

Georgia State University

A Paper Presented at the Annual Meeting of the

National Council on Measurement in Education

Boston, MA, 1990

Abstract

This study examined evidence of construct validity of the revised Special Education-Mental Handicaps Georgia Teacher Certification Test (MH-TCT) using hierarchical confirmatory factor analysis and LISREL VI. The primary objective was to determine if first and second order factors corresponded to item/objective/test relationships defined by judgmental/consensual content validation studies. As sample size was small, findings were tentative. LISREL was used to examine the factor structure of the entire test as well as to identify test items that were not psychometrically sound. Although six subdimensions were identified which corresponded to the six test objectives, first-order factors corresponding to these were so highly correlated that a hierarchical model was rejected in favor of a unidimensional model. A unidimensional model was invariant across samples and time. Results suggested that the LISREL confirmatory maximum likelihood factor model provided a powerful tool for item analysis and test validation. A major drawback was the amount of computer run time needed to do maximum likelihood factor analysis.

Thanks are extended to Dr. Richard M. Jaeger and Dr. William A. Mehrens who serve on the Georgia Technical Advisory Committee and Dr. Madelaine C. Ramey of Seattle Public Schools for advice and recommendations regarding this study.

A Validation Study of a Georgia Teacher Certification Test Using Confirmatory Factor Analysis

This study was undertaken by the Georgia Department of Education to determine if test development procedures designed to provide evidence of conte.it validity of the Georgia Teacher Certification Tests were supported by studies of construct validi. ,. The new Standards (APA, AERA, & NCME, 1985) as well as the third edition of Educational Measurement (Linn, 1989) define test validity as a unitary concept which should be examined from a number of perspectives. Using this substantive approach, initial judgmental/consensual procedures, used to make decisions about domain definition, test specifications, and item selection are substantiated or revised based on empirical data. Traditionally, factor analysis has been a major method used to study construct validity because it provides a succinct method of identifying convergent and divergent patterns in the data as well as irrelevant test method variance resulting from psychometric flaws in items (Messick, 1989). Until recently, only exploratory factor analysis could be conducted using commercially available computer programs. LISREL VI (Joreskog & Sorbom, 1986) allows the researcher to conduct confirmatory as well as exploratory factor analysis, thereby providing the means to test hypotheses regarding substantive relationships in the data.

Several recent studies have used confirmatory maximum likelihood factor analysis to examine construct validity of a measure of the structure-of-intellect model (Dunmier, Michael, & Hocever, 1988) and the Stanford-Binet Fourth Edition (Keith, Cool, Novak, & White, 1988; Keith, Cool, Novak, White, & Pottebaum, 1988). These authors concluded that confirmatory factor analysis provided a much stronger indication of the underlying structure of a test than exploratory factor analysis because much less subjectivity is involved when the factor structure is specified in advance. Harris (1985) stated that the increasing use of confirmatory factor analysis is the most promising development in factor analysis in recent years, as problems related to interpretation of factors based on factor loadings as well as decisions regarding the number of factors to be extracted are avoided if variables are assigned to factors on the basis of a priori specified relations to underlying conceptual variables. On the other hand, Paunonen (1987) examined a multidimensional group of items using several types of factor analysis incluoing confirmatory maximum likelihood method and concluded that multiple group factor analysis provided few practical benefits over traditional item-total correlational procedures.

This study examined evidence of construct validity of the revised Special Education-Mental Handicaps Georgia Teacher Certification Test (MH-TCT) using hierarchical maximum likelihood confirmatory factor analysis and LISREL VI (Joreskog & Sorbom, 1986). The primary objective was to determine if first and second order factors corresponded to item/objective/test relationships defined by judgmental/consensual content validation studies. The construct measured by the test was defined for this study as being equivalent to the content domain of knowledge and skills needed to teach students with mental handicaps in public schools in Georgia. The domain and objectives of the MH-TCT had been defined by content experts during the domain definition process. Detailed test specifications were developed in accordance with the defined domain. After items were written, content experts reviewed items to ensure item-objective match. For this study, a hierarchical factor model was conceptualized as a two-tier structure in which a single second-order factor corresponded to the overall test domain and six correlated first-order factors corresponded to the six MH-TCT objectives. Correlations among first-order factors were expected to be fairly high because of substantive item selection criteria which excluded items with low *r*-point biserial item-total test correlation coefficients.

A second objective of the inquiry investigated the possibility of using maximum likelihood confirmatory factor analysis and LISREL VI as a tool for test scale development. At the item level, a number of alternatives to classical test statistics have been studied. Although item response theory (IRT) provides a powerful means of scale development, the various models subsumed under that theory generally require strong assumptions, including unidimensionality, which is not a requirement for the TCTs. Hsu and Yu (1989) summarized studies that focused on the impact of violations of the unidimensionality assumption on IRT models and concluded that although IRT models are robust to moderate amounts of multidimensionality, violations could seriously affect parameter estimates. Confirmatory maximum likelihood factor analysis has two major requirements for calculation of the chi-square test, multivariate normality and large sample size. However, the number of common and unique factors and the correlations among factors are subject only to restrictions imposed based on substantive interpretation of relationships in the data. As the confirmatory factor analysis model of LISREL VI allows the researcher to specify parameters for a factor model so that the configuration of factors as well as the correlation among factors are specified in advance, items could be assigned to factors based on substantive criteria. It was hypothesized that goodness of fit statistics for the maximum likelihood confirmatory factor analysis model as well as item-level factor loadings, modification indices, standard errors, and _t_-values which were outside of prespecified acceptable ranges could be used to identify flawed field test items.

A final objective compared performance of items measuring memorization of factual information with items measuring higher level cognitive behaviors. Items were classified by content and cognitive process in accordance with test specifications developed during the domain definition process. Cognitive behaviors were defined using Bloom's _Taxonomy of Educational Objectives_ (Bloom, Engelhart, Furst, Hill, & Krathwohl, 1956). Generally, Bloom's work has been interpreted to mean that items measuring thought processes at higher levels of the hierarchy are more difficult than those at lower levels which measure the same content, as they subsume lower behaviors. However, few studies exist which systematically explore the complex interrelationships of content to cognitive process as reflected in performance on test items. The MH-TCT test specifications as well as cognitive learning theory (Snow & Lohman, 1988) predict multiple sources of item difficulty which include content difficulty as well as type and complexity of cognitive processing required which was expected to be reflected in the factor structure of the test.

Instruments Used

In 1986, the Georgia Department of Education contracted with Georgia Assessment Project of Georgia State University to revise and/or develop the Georgia Teacher Certification Tests. Under Georgia law, candidates who apply for state certification in education must pass a screening test in the certification area in which they are applying. Currently, TCTs exist for 28 content areas. The TCTs are paper and pencil tests designed to measure knowledge and skills judged necessary for basic functioning as a teacher or administrator. Extensive discussion of the TCT revision process is available in the TCT Technical Manual (Georgia Assessment Project, in press).

The Special Education-Mental Handicaps TCT (MH-TCT) is a multiple choice test developed by Georgia special education teachers. For each TCT, content specialists (practitioners and teacher trainers) in respective subject areas were trained in various aspects of test development by Georgia Assessment Project of Georgia State University. Item writer guides which provide detailed test specifications were developed by GAP under advisement of Georgia teachers and were used throughout the test development process. Objectives and assessment characteristics are described in Appendix A.

Test development procedures followed prescribed judgmental/consensual methods outlined in detail in the Georgia Teacher Certification Test Technical Manual developed by Georgia Assessment Project (in press). These procedures include administration of a job analysis questionnaire, domain definition/ objective validation, item writer training and item writing, item content and bias review, final item/objective match, field test and review of item/test statistics, and standard setting. A chart outlining the test development process is attached (Appendix B).

The new form of the MH-TCT consisted of 143 multiple choice items. Of those, 92 were actual scorable items which had been field tested. To be included on the test as a scorable item, each had to pass a content/bias review by subject matter experts and had to have item statistics within acceptable range (*r*-point biserial significantly greater that zero), according a formula for approximating the standard deviation of correlation coefficients (Crocker & Algina, 1986, p. 34).

Method

Data from four administrations of the MH-TCT were used in the analyses. Subjects were 287 candidates for certification in the mental handicaps subarea of special education. The number of examinees was reduced considerably from estimates based on previous years. Analyses were either modified or results were qualified because of the reduced sample size.

Hierarchical maximum likelihood factor analysis using LISREL VI (Joreskog & Sorbom, 1986) was used to determine if hypothesized combinations of factors could be identified which correspond to the item/objective/test relationships identified during the test development process and outlined in the test specifications. The covariance rather than correlation matrix was used in each of the following analyses unless otherwise indicated.

The hierarchical confirmatory maximum likelihood factor analysis model (Joreskog & Sorbom, 1986, p. I.10) was used to test the hypothesis that a single second-order factor corresponding to the test domain and six first-order correlated factors corresponding to the test objectives would imply a correlation matrix which is the population correlation matrix (Harris, 1985). The small number of subjects relative to the large number of items resulted in the potential for a highly unstable factor solution. Further, computer memory would not accommodate the entire set of 143 items. Consequently, five scorable items per objective were selected for initial study, following research done by Paunonen (1987) on confirmatory factor analysis using LISREL. Parameters for the model were specified so that all items not related to a factor had factor coefficients which were fixed to be equal to zero. Factor variances were fixed at unity. All other parameters (unique variance and first-order factor correlations) were allowed to vary.

Output requested included first and second order factor loadings, modification indices, standard errors, and t-values (defined as a parameter divided by its standard error).

As the small sample size relative to the large number of variables resulted in a potentially unstable factor solution, a second hierarchical maximum likelihood factor analysis was conducted in which the covariance matrix for the six test objectives was input for analysis. An advantage of this analysis was that it provided a model which took all the scorable items into account, providing a cross-check on the first model while reducing the number of parameters in order to increase factor stability. Based on information obtained from the first analysis, a unidimensional model was specified and compared with a single-factor model with two correlated subfactors. An additional analysis compared fit statistics before and after a priori restrictions on the factor structures were dropped for the hierarchical model. For the unidimensional model, parameters were specified so that all objectives loaded on a single factor with factor variance set at unity. For the single factor hierarchical model with two correlated subfactors, factor coefficients not substantively related to one or the two factors were set at zero and factor variances were fixed at unity. Unique variances and first-order factor correlations were allowed to vary, as with the previous model. For the hierarchical model, objectives were assigned to one of two subfactors based on substantive similarities or differences. The three objectives that measured developmental theory, state and federal law, and IEP development were assigned to one factor and the three objectives that measured curriculum/instruction, diagnosis and assessment, and classroom management were assigned to the other. Objectives assigned to the first factor were made up predominantly of comprehension or application level items while objectives assigned to the second contained more knowledge level items.

A third exploratory analysis was conducted to determine if factor structures specified in the second analysis above were invariant across samples of examinees and across time. Data from four test administrations were combined so that 125 subjects from the first and third test administrations comprised group 1 and 125 subjects from the second and fourth administrations comprised group 2. Using the six test objectives as first order factor loadings as specified in analyses in the previous paragraph for (a) the unidimensional model and (b) the single-factor hierarchical subfactor model, five hypotheses involving equality of factor structures were tested and compared in accordance with the example in LISREL VI (Joreskog & Sorbom, 1986, pp. V.3-V.12). Hypotheses regarding equivalence in the two samples were as follows:

1. Hypothesis 3a - Covariance matrices are equivalent.
2. Hypothesis 3b - The number of factors is equivalent.
3. Hypothesis 3c - First-order factor loadings are invariant.
4. Hypothesis 3d - Unique variance is invariant.
5. Hypothesis 3e - Second-order factor matrices are equivalent.

A fourth analysis using maximum likelihood confirmatory factor analysis (Joreskog & Sorbom, 1986, p. I.9) was conducted to examine submatrices consisting of items assigned to a single objective to determine if item-objective fit was adequate according to fit statistics for the entire model. In addition, item level criteria were applied, including non-trivial factor loadings (.10 or greater, after Paunonen, 1987) and _t_-values significantly greater than zero. Two analyses per objective were run. The total fit statistics obtained for an objective with only scorable items were compared with fit obtained when new field test items were included. Factor loadings, standard errors, and *t*-values obtained in these analyses were compared with *r*-point biserial correlations to determine if the LISREL confirmatory factor analysis model increased the efficiency of item selection.

Results

As sample size was half that expected, results were interpreted with caution, especially those involving a large number of parameters compared to the number of examinees. Adequacy of sample size is a function of the number of parameters specified. Nunnally (1978) recommends using 10 times the number of independent variables in the model as a lower bound for acceptable sample size. Consequently, analyses at the objective level were considered more stable than those at the item level. Results of analyses were as follows:

Analysis 1. Hierarchical maximum likelihood confirmatory factor analysis of the 30-item, 6-factor item/objective/test matrix resulted in a phi matrix which was not positive definite because of extremely high correlations among first-order factors. Otherwise, all parameters were within acceptable bounds. Each item loaded non-trivially on the factor to which it was assigned (greater than or equal to .10, after Paunonen, 1987), according to maximum likelihood estimates. The goodness of fit index, a measure of the amount of variance and covariance jointly accounted for by the model, was .904. The root mean square residual, which is a measure of the average residual variance and covariances, was .055. The chi-square test was significant. However, Joreskog and Sorbom (1986) caution against rejection of a model based on the chi-square test, as it is frequently significant with large samples, even when data-model fit is good. Overall results for the entire model suggested that although the first-order specifications of the six factor model resulted in a good fit (see Table 1), the extremely high correlation among first-order factors resulted in the need for additional exploratory analyses with fewer factors.

Exploratory factor analysis yielded one factor which accounted for 52 percent of test variance, using Kaiser's criteria that factors must have eigenvalues greater than or equal to one for inclusion. Consequently, two single-factor models were compared using maximum likelihood factor analysis. The first was a unidimensional model in which all items loaded on a single factor. The second was a single-factor model with two subfactors which corresponded to substantive relationships in the data. Model specifications were set so that items which measured factual knowledge of developmental theory, state and federal law, and state and federal requirements (from objectives 2, 4, and 5) loaded on one subfactor and items which measured knowledge of diagnosis and assessment, teaching strategies, and behavior management (from objectives 1, 3, and 6) loaded on a second subfactor.

Results, found in Table 1, indicated little difference in fit for either model. Although the chi-square test (a likelihood ratio test statistic for testing the probability of obtaining a chi-square value larger than the value obtained given the fact that the specified model is accurate) was significant for both models, Joreskog and Sorbom (1986) caution against rejection of a factor model based on that criteria, as chi-square is often significant with large samples, even when good fit to the data are found. For the current analysis, the chi-square test, goodness-of-fit index, adjusted goodness-of-fit index, and root mean square residual indicated slightly better fit for the model with two subfactors than for the unidimensional model. However, the extremely high correlation of first-order factors yielded an element in the first-order factor correlation matrix that was not identified. Examination of the phi matrix of the hierarchical model indicated a correlation of .91 of the two first-order factors. Although significant improvement in fit could be observed when factor loadings were estimated after dropping a priori restrictions, the extremely high correlation of factors again resulted in an error message that an element of the phi matrix was not identified. Consequently, both hypotheses regarding the hierarchical model (a priori restrictions provide the best fit versus no a priori restrictions provide the best fit) were rejected in favor of the unidimensional model.

Analysis 2. Results of confirmatory maximum likelihood factor analysis using test objective total scores as first-order factor loadings are found in Table 2. Following the previous analyses, two models were compared. The first was a unidimensional model in which all objectives loaded on the same factor. The second was an hierarchical model in which first order factor loadings corresponded to either knowledge or comprehension and application taxonomic levels specified in the previous analysis. While the latter model exhibited somewhat better fit to the data, the extremely high correlation of first-order factors resulted in an element in the second-order correlation matrix that was not identified, consistent with results of Analysis 1.

Analysis 3. Results of tests of equality of factor structures are found in Table 3. Overall fit statistics were compared for the two factor models described in analysis 2. These analyses were done at the objective rather than item level for two reasons. First, the objective-level model contained information from the full set of scorable items rather than a sample. Second, as the number per sample was reduced to 125 subjects per group in order to compare factor structures of two samples, only the objective-level analysis came anywhere near acceptable criteria for the number of cases needed for a stable multivariate analysis relative to the number of model parameters. Analyses were conducted in stepwise fashion after recommendations of Joreskog and Sorbom (1979). Although the unidimensional model was thought to be the appropriate model, the second model described in Analysis 2, the single factor, two-subfactor model, was also examined, as the pattern in that data was consistent with substantive theory. Hypothesis 3a, covariance matrices are equivalent for the two samples, was accepted. Hypothesis 3b, the number of factors is equivalent, held for both the unidimensional and hierarchical factor models. Hypotheses 3c and 3d, invariance of first-order common and unique factors, also held for both. Hypothesis 3e, second-order factor matrices were equivalent, held for the hierarchical model. Although the single factor hierarchical model showed slightly better fit than the unidimensional model, the extremely high intercorrelation of first-order factors found in Analysis 2 precluded acceptance of that model, as an element of the phi matrix was not identified. Clearly additional study witn a larger number of subjects and item-level data is needed to determine if that pattern is an anomaly of the current small sample or a true pa tern in the data.

Analysis 4. Analysis of six submatrices made up of all item level data for each of the six objectives examined the item/objective relationship. Item/objective fit statistics, first-order factor loadings, standard errors, and *t*-values were examined to determine if items were psychometrically sound.

As the overwhelming evidence of previous analyses suggested that the MH-TCT was essentially unidimensional, six unidimensional models were specified at the item level for the six test objectives. Table 4 presents item/objective fit statistics for each objective for (a) initial scorable items only and (b) scorable and field test items.

As scorable items had been selected initially because they met specific content and measurement criteria, good fit was expected and was actually found. All measures of goodness of fit reflected the impact of field test items which did not fit the model well. Tables 4 provides measures of goodness of fit for each objective with and without field test items. Scorable items conformed to prespecified criteria. Factor loading were non-trivial (.10 or greater, after Paunonen, 1987); standard errors were not significantly greater that zero, according a formula for approximating the standard deviation of correlation coefficients, given by Crocker & Algina (1986, p. 34); and _t_-values were significantly greater than zero. Modification indices did not function with a single factor. Each objective had two or more field test items which did not fit the above criteria. The fit statistics for scorable and field test items indicated that for all objectives, poorly fitting items did lower the overall fit of the model as reflected in the chi-square test and in the other fit statistics.

Further item-level analyses compared results of factor analyses with classical item statistics. The item- total test _r_-point biserial correlation coefficient was compared with LISREL first-order factor loading and _t_-values to determine if the same field test items were identified as defective across indices. Item level data from the six item-objective analyses above were used to provide factor loadings and _t_-values. LISREL confirmatory factor analysis results identified the same items as either psychometrically sound or defective for 62 percent of the items. The LISREL factor loading was out of bounds (smaller than .10) while the _r_-point biserial was in bounds (significantly greater than zero) for 30 percent of the items. The _r_-point biserial value was out of bounds while the LISREL factor analysis results were in bounds for 3 percent of the items. The magnitude of the first-order item level factor loadings decreased as the number of items per objective increased.

Conclusions

Overall results suggest that the MH-TCT is essentially a unidimensional measure. Six subfactors were identified using hierarchical confirmatory maximum likelihood factor analysis and LISREL VI which corre:ıonded to the six MH-TCT objectives identified during judgmental/consensual validation process. However, first-order factors were so highly correlated that the model was tentatively rejected in favor of a unidimensional model. It appeared that although six factors could be identified which were consistent with test specifications, they were so closely correlated that they appeared to be different manifestations of a single entity. Based on those initial findings, two single-factor models were hypothesized and compared at the item and a' the objective levels, based on substantive theory that the total test measured a single domain, knowledge and skills necessary to teach mentally handicapped students in Georgia public schools. The first was a unidimensional model. The second was a single-factor hierarchical model in which parameter specifications were fixed so that two correlated subfactors were set which corresponded to objectives measuring factual information versus those measuring application of knowledge. Objectives 2, 4, and 5, which covered developmental theory and knowledge of the law corresponded most closely to the first factor and objectives 1, 3, and 6, which covered diagnosis and assessment, curriculum/instruction, and behavior management corresponded to the second. Both models showed extremely good fit at the item and at the objective levels, as reflected in overall goodness of fit statistics. Joreskog and Sorbom (1986) recommend using the root mean square residual, defined as the average of the residual variances and covariances, to compare fit of two different models for the same data. Although the latter hierarchical model showed slightly better fit using this criterion, one of the elements of the phi matrix was not identified because of extremely high correlations among first order factors, which led to tentative rejection of that model in favor of the unidimensional model.

High correlations for the six test objectives were not unexpected, as test items are selected based on item-total test correlations. Some research suggests that the LISREL VI confirmatory maximum likelihood method of factor analysis may yield lower targeted factor loadings and higher factor correlations than so other factor analytic procedures (Paunonen, 1987). Had criteria for inclusion been applied to item-objective rather than item-total test correlations it would have been possible although not probable for factors corresponding to objectives to be orthogonal. However, since the standard setting process that established the pass/fail score for the MH-TCT was done in relation to the total test, it appeared prudent to apply item-total test rather than item-objective criteria, in accordance with classical test theory regarding test reliability and the standard error of measurement.

The five tests for equality of factor structures yielded results which suggest that a unidimensional factor structure is invariant across samples and time. These findings substantiated the hypothesis that a unidimensional factor structure would reproduce the population correlation matrix implied by the factor loadings. Using test objectives as first-order factor loadings, both the unidimensional model and the hierarchical model described in the previous paragraph showed good data-model fit. Although substantive theory regarding systematic differences in performance of knowledge-level versus comprehension or application-level items as well as improved fit statistics supported the hierarchical model, the small sample size and high correlations among factors suggest that judgment should be withheld beyond a unidimensional interpretation pending further research on a larger number of subjects.

LISREL VI and confirmatory maximum likelihood factor analysis appeared to provide a useful supplementary tool for item analysis. In comparison to traditional exploratory factor analysis models, the current LISREL model provides total control over model specifications. This ability to develop a priori factor models and to test those models statistically for fit represents a major improvement over exploratory factor analysis. In addition to fit statistics for the entire model, modification indices, standard errors, and _t_-values provide tools for identification of individual items which are not functioning properly. Of note, results of the factor analyses were in agreement with classical item statistics approximately two-thirds of the time when used to identify psychometrically sound field test items. As factor loadings are affected by scale length as well as item-scale fit, it is not clear whether the field test items rejected by only the factor analysis model were actually faulty. Additional review by content experts is needed to resolve this issue.

A major drawback of maximum likelihood factor analysis noted by Harris

(1985) and Paunonen (1987) is the large amount of computer memory needed to do either exploratory or confirmatory maximum likelihood factor analysis. Analysis of item-level data for tests of even moderate lengths like the MH-TCT become unwieldy. Maximum likelihood factor analysis is expensive to run on mainframe computers compared to classical item statistics as well as more conventional methods of performing factor analysis or principal components analysis. However, this limitation can be at least partially overcome by analysis of test submatrices and allocation of additional computer memory.

Of primary importance, this preliminary study suggests that confirmatory factor analysis can be a powerful tool when used to examine evidence of test validity, as the researcher can test hypotheses about item/subtest/total test relationships against a priori specifications based on results of judgmental/ consensual procedures used to obtain evidence of content validity. Although small sample size and limited computer memory precluded analysis of an item-level content-by-process matrix for the current study, the LISREL VI confirmatory factor analysis program provides the degree of control of model specifications necessary to conduct future studies of that nature.

References

Arter, J. A., & Salmon, J. R. (1987). Assessing higher order thinking skills: A consumer's guide. Portland, OR: Northwest Regional Educational Laboratory.

8loom, 8. S.; Englehart, M. 8.; Furst, E. J.; Hill, W. H.; & Krathwohl, D. R. (1956). Taxonomy of educational objectives. The classification of educational goals. Handbook I: Cognitive domain. New York: Longmans Green.

Dunmire, P. A.; Michael, W. 8.; & Hocever, D. J. (1988). Confirmatory maximum likelihood factor analysis as applied to measures of social intelligence within the structures-of-intellect model. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, Louisiana.

Georgia Assessment Project (in print, 1989). A technical manual for the revised Georgia Teacher Certification Tests. Atlanta, GA: Georgia State University.

Harris, R. J. (1985). A Primer of Multivariate Statistics. New York: Academic Press, Inc.

Hsu, T. & Yu, L. (1989). Using computers to analyze item response data, Educational Measurement: Issues and Practice,8(3).

Joreskog, K. G., & Sorbom, D. (1984). LISREL VI: Analysis of linear structural relationships by maximum likelihood, instrumental variables, and least squares methods: User's guide. Mooresville, IN: Scientific Software.

Joreskog, K. G., & Sorbom, D. (1979). Advances in Factor Analysis and Structural Equation Models. New York: University Press of America.

Keith, T. Z.; Cool, V. A.; Novak, C. G.; & White, L. J. (1988). Hierarchical confirmatory analysis of the Stanford-Binet Fourth Edition: Testing the theory-test match. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, Louisiana.

Keith, T. Z.; Cool, V. A.; Novak, C. G.; White, L. J.; & Pottebaum, S. M. (1988). Confirmatory factor analysis of the Stanford-Binet Fourth Edition: Testing the theory-test match. Journal of School Psychology, 26, 253-274.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational Measurement (3th ed.). New York: Macmillan Publishing Company.

Nunnally, J. C. (1978). Psychometric Theory. (2nd. ed.). New York: McGraw-Hill.

Paunonen, S. V. (1987). Test construction and targeted factor solutions derived by multiple group and Procrustes methods. Multivariate Behavioral Research,22,437-455.

Snow, R. E., & Lohman, D. F. (1989). Implication of cogitive psychology for educational measurement. In R. L. Linn (Ed.), Educational Measurement (3th ed.). New York: Macmillan Publishing Company.

Table 1

Hierarchical Confirmatory Maximum Likelihood Factor Analysis

Comparison of Measures of Goodness of Fit for Three Models

Using Item-Level Data

	Measures of Goodness of Fit		
	6 Factors Hierarchical*	1 Factor	1 Factor Hierarchical**
Chi-square	473.13(p=.002)	554.31(p=.000)	545.23(p=.000)
	(df=390)	(df=405)	(df=404)
Goodness of fit index	.904	.888	.890
Adjusted goodness of fit index	.886	.872	.874
Root mean square residual	.055	.055	.055

n=287

*The phi matrix was not positive definite.

**The model was not identified.

Table 2

Factor Loadings and Fit Statistics for Objective-level Data:

Confirmatory Maximum Likelihood Factor Analysis

	1 Factor	1 Factor Hierarchical	
First-order factor loadings:	f1	f1	f2
Assessment	.791	.877	0
Characteristics	.831	0	.900
Curriculum/Instruction	.825	.916	0
State/Federal Law	.590	0	.619
IEP Development	.666	0	.715
Classroom Management	.549	.615	0
First-order factor correlation		.91*	
Fit statistics			
Chi-square	21.32(p=.011) (df=9)	10.43(p=.108) (df=6)	
Goodness of fit index	.973	.988	
Adjusted goodness of fit index	.938	.958	
Root mean square residual	.030	.021	
Coefficient of determination	.880	.925	

*Output indicated that this element may not be identified.

Table 3

Stabi̇lity of Factor Structures of the MH-TCT:

Tests of Five Hypotheses Using Objective-Level Data

	1 factor	1 factor Hierarchical
Hypothesis A:		
Chi-square	21.02(p=.458)	21.02(p=.458)
Goodness of fit index	.978	.978
Root mean square residual	.390	.390
Hypothesis 8:		
Chi-square	29.12(p=.023)	17.15(p=.144)
Goodness of fit index	.971	.993
Root mean square residual	.189	.098
Hypothesis C:		
Chi-square	34.46(p=044)	21.58(p=.251)
Goodness of fit index	.964	.986
Root mean square residual	.339	.247

(Continued on next page)

Table 3(Cont.)

	1 factor	1 factor Hierarchical
Hypothesis D:		
Chi-square	42.53(p=.039)	28.69(p=.232)
Goodness of fit index	.958	.979
Root mean square residual	.364	.251
Hypothesis E:		
Chi-square		31.65(p=.245)
Goodness of fit index		.979
Root mean square residual		.382

Hypothesis A = covariance matrices are equivalent

Hypothesis B = the number of factors is the same in both groups

Hypothesis C = common factors are invariant

Hypothesis D = unique factors are invariant

Hypothesis E = second-order factor structures are invariant

Table 4

Item Level Fit Statistics by Objective:

	Scorable Items	Scorable + Field Test Items
Objective 1:		
Chi-square	99.40(p=.234)	262.33(p=.070)
Goodness of fit index	.957	.927
Adjusted goodness of fit index	.943	.912
Root mean square residual	.047	.051
Objective 2:		
Chi-square	115.03(p=.032)	455.57(p=.000)
Goodness of fit index	.963	.893
Adjusted goodness of fit index	.945	.874
Root mean square residual	.010	.011

(Continued on next page)

Table 4 (Cont.)

	Scorable Items	Scorable + Field Test Items
Objective 3:		
Chi-square	494.56(p=.000)	1003.03(p=.000)
Goodness of fit index	.893	.856
Adjusted goodness of fit index	.877	.841
Root mean square residual	.011	.011
Objective 4:		
Chi-square	23.06(p=.286)	133.79(p=.026)
Goodness of fit index	.980	.945
Adjusted goodness of fit index	.965	.929
Root mean square residual	.010	.011
Objective 5:		
Chi-square	44.50(p=.130)	80.25(p=.096)
Goodness of fit index	.972	.960
Adjusted goodness of fit index	.955	.944
Root mean square residual	.008	.009

(Continued on next page)

Table 4 (Cont.)

	Scorable Items	Scorable + Field Test Items
Objective 6:		
Chi-square	71.39(p=.247)	175.23(p=.011)
Goodness of fit index	.961	.938
Adjusted goodness of fit index	.946	.921
Root . ean square residual	.009	.010

Field 15: Mental Handicaps
Objectives

Objective 01: The educator identifies and applies **principles of student assessment procedures** in the context of the instructional environment and academic materials. This objective accounts for approximately 15-20 percent of the items on the test.

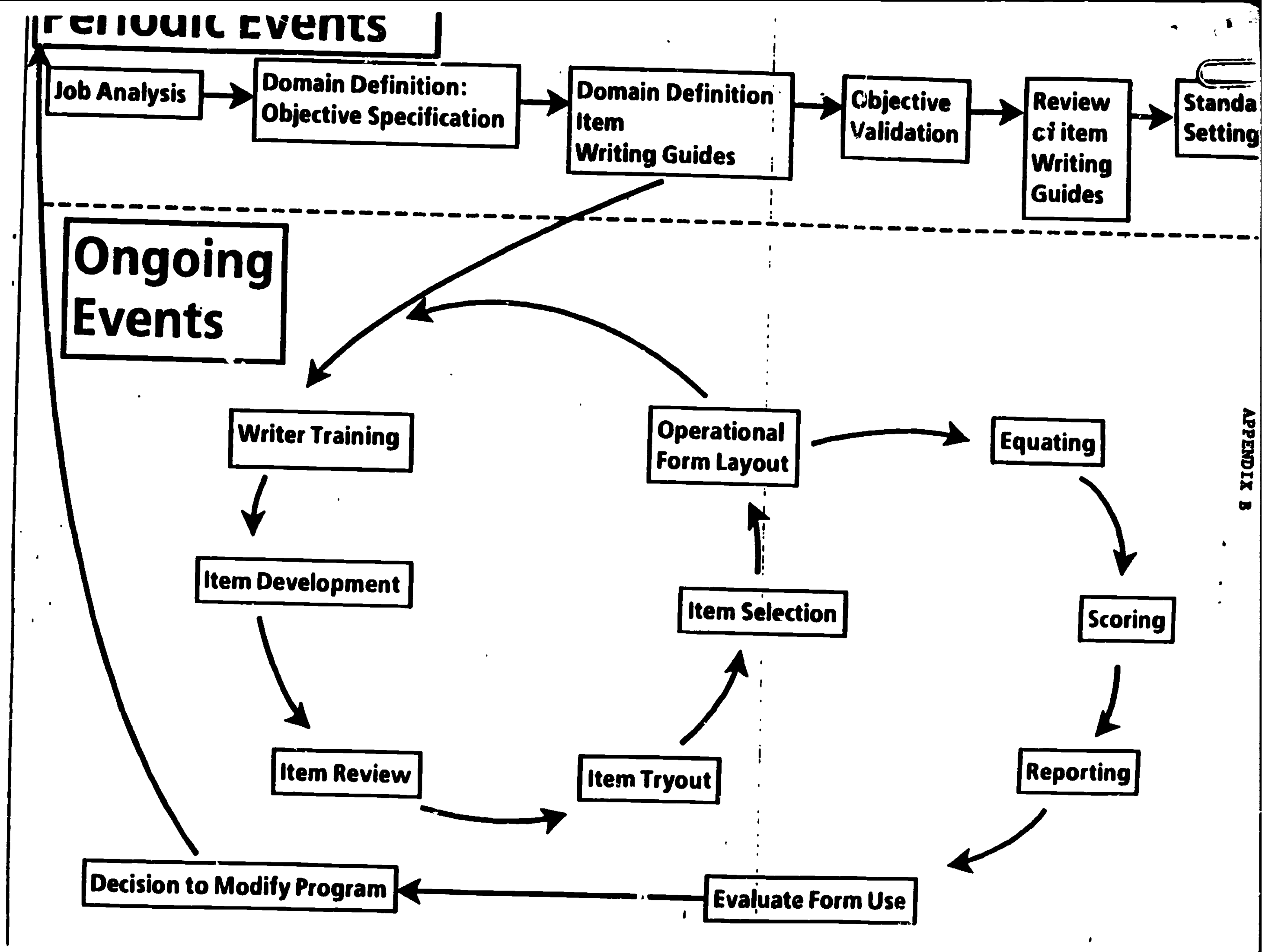
Objective 02: The educator identifies **characteristics of individuals with mental handicaps and the implications of these handicaps for learning** in the context of the instructional environment and academic materials. This objective accounts for approximately 15-20 percent of the items on the test.

Objective 03: The educator identifies **curriculum content and instructional strategies for individuals with mental handicaps** in the context of the instructional environment and academic materials. This objective accounts for approximately 20-25 percent of the items on the test.

Objective 04: The educator identifies **principles of state and federal laws, rules, regulations and policies that apply to special education** in the context of the instructional environment and academic materials. This objective accounts for approximately 10-15 percent of the items on the test.

Objective 05: The educator identifies **principles related to the development and implementation of the individualized Education Program (IEP)** in the context of the instructional environment. This objective accounts for approximately 15-20 percent of the items on the test.

Objective 06: The educator identifies **principles of classroom management** in the context of the instructional environment and academic materials. This objective accounts for approximately 15-20 percent of the items on the test.



Periodic Events

Job Analysis

Domain Definition: Objective Specification

Domain Definition Item Writing Guides

Objective Validation

Review of item Writing Guides

Standa Setting

Ongoing Events

Writer Training

Item Development

Item Review

Item Tryout

Item Selection

Operational Form Layout

Equating

Scoring

Reporting

Evaluate Form Use

Decision to Modify Program

APPENDIX B