ED 431 014 TM 029 846

AUTHOR Li, Yuan H.; Ford, Valeria; Tompkins, Leroy J.
TITLE The Construct Validity of a Performance-Based Assessment Program.
PUB DATE 1999-04-00
NOTE 44p.; Paper presented at the Annual Meeting of the American Educational Research Association (Montreal, Quebec, Canada, April 19-23, 1999).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Construct Validity; Correlation; Elementary Education; Elementary School Students; Grade 3; Grade 5; Multiple Choice Tests; *Performance Based Assessment; State Programs; *Structural Equation Models; Testing Programs
IDENTIFIERS Comprehensive Tests of Basic Skills; *Maryland School Performance Assessment Program

ABSTRACT

The purpose of this study was to examine the construct validity of a performance assessment program, the Maryland School Performance Assessment Program (MSPAP). Based on analyses of the longitudinal associations of Grade 5 MSPAP data in 1996 with Grade 3 MSPAP data in 1994, the following hypothesis was examined: the unattentuated correlation or the group-mean correlation between two similar measures of the same content area is higher than its correlations with different content areas. This hypothesis was not supported. In addition, the results analyzed by structural equation modeling (SEM) of this longitudinal correlation matrix reveal that the SEM model specified by the MSPAP six latent traits was unable to capture the underlying information of this data. Extra factors, such as a general ability and an assessment method effect, may need to be considered for better fitting data. SEM was performed on the multitrait-multimethod correlation data, and the traits of Reading and Mathematics were assessed by MSPAP and the Comprehensive Test of Basic Skills (CTBS). The trait effects of MSPAP reading and CTBS mathematics application may be attenuated by the method effects of the performance-based assessment and the multiple-choice assessment, respectively. (Contains 6 figures, 7 tables, and 19 references.) (Author/SLD)

The Construct Validity of a Performance-based Assessment Program

Yuan H. Li, Valeria Ford, Leroy J. Tompkins

Prince George's County Public Schools, Maryland

Paper presented at the annual meeting of the American Educational Research Association, April, 19-23, Montreal, Canada

The Construct Validity of a Performance-based Assessment Program*

Abstract

The purpose of this study is to examine the construct validity of a performance assessment program, the Maryland School Performance Assessment Program (MSPAP).

Based on analyses of the longitudinal associations of Grade 5 MSPAP data in 1996 with Grade 3 MSPAP data in 1994, the following hypothesis was examined: the unattenuated correlation or the group-mean correlation between two similar measures of the same content area is higher than its correlations with different content areas. This hypothesis was not attained. In addition, the results analyzed by structural equation modeling (SEM) to this longitudinal correlation matrix reveal that the SEM model specified by the MSPAP six latent traits was unable to capture the underlying information of this data. Extra factors, such as, a general ability and an assessment method effect, may need to be considered for better fitting the data.

SEM was performed on the multitrait-multimethod correlation data, where the traits of Reading and Math was assessed by MSPAP and CTBS (the Comprehensive Test of Basic Skills). The trait effects of MSPAP reading and CTBS mathematics application may be attenuated by the method effects of the performance-based assessment and the multiple-choice assessment, respectively.

Key Words: Construct Validity; Reliability; Performance-based Assessment;
Structural Equation Modeling

* The viewpoints made in this study reflect the authors' opinions rather than that of the Prince George's County Public Schools, Maryland.

I. Introduction

The Goals 2000, Educate America Act passed in 1994, specifies that high learning standards and innovative forms of assessments should be used as the chief means to ensure that educational reform is on the right track. School systems are required to look beyond the traditional method of multiple-choice testing for better forms of assessments in evaluating students' achievements. Of the many innovative forms of assessments suggested, performance-based assessments have been widely adopted. They usually require students to construct a variety of responses to test items or tasks that are similar to classroom instructional activities and to those used in real life.

Proponents of the performance-based assessment are of the opinion that all of the real or perceived shortcomings of traditional assessments would be remedied by this transition. Resulting improvements include more valid measures of student performance, elimination or reduction in bias or perceived bias in traditional assessments, etc. In light of some of the issues that remain unresolved, more psychometric questions were addressed (literature review by Green, 1995). For example, nonstandardized test formats or testing procedures, difficulty in maintaining the test specifications, inconsistent scoring rubrics, differences in the raters' severity or leniency in scoring, violating underlying principles for modeling test data or equating tests, etc. may render the scores as being biased as well as not being comparable from one year to the next.

In a school improvement instructional model that includes a high stakes testing program, data resulting from a performance assessment model MUST provide building managers and teachers with directions regarding the strengths and weaknesses in their instructional programs. The extent to which the data does or does not accurately provide this direction is an indicator of the validity in the assessments using such a model. Not withstanding concerns pertaining to accountability from an administrative perspective, the extent to which the scores accurately reflect where there are strengths and deficiencies in the instructional program is absolutely critical. This is because time, effort and substantial resources must be allocated to address those areas which are deficient. In a data-driven instructional program, the reliability and validity of the data are paramount if schools are to successfully attain the prescribed standards.

The purpose of this study is to examine the construct validity of a performance assessment program by analyzing the performance-based test data set in one school district. We expect the results from this study to provide valuable assistance to other similar performance-based assessment programs.

Background of a Performance-based Assessment Program

Since 1991, the Maryland State Department of Education (MSDE) has implemented the annual Maryland School Performance Assessment Program (MSPAP) for grades 3, 5, and 8 in all of its public schools. The MSPAP assessments consist of six content areas, Reading (RDSS), Writing (WRSS), Language Usage (LSS), Mathematics (MSS), Social Studies (SSSS) and Science (SCSS). MSPAP was an innovative performance-based assessment. The primary focus of the information provided from MSPAP assessments is school performance rather than individual student performance because of the design of the MSPAP's test and its sampling design. Performance on the MSPAP has been used to evaluate whether schools meet a satisfactory standard that was set by the State Department of Education. Schools that consistently do not meet the standard may be managed by an outside organization if their MSPAP performance does not improve. It becomes apparent that with MSPAP provided accountability, the score report is very important to test practitioners as well as school authorities.

MSPAP test items (tasks) are integrated both within a content area and across content areas so that students have an opportunity to integrate information they have learned (Maryland State Department of Education, 1996). To cover the required breadth of learning outcomes in limited testing time, three non-parallel test forms per content area were developed and randomly assigned to students within a school. 'Non-parallel' means that the test tasks of the three test forms are not completely created from the same domains (a group of learning outcomes). An equating design was used for tracking schools' yearly improvement. Three steps are taken to equate MSPAP scales between two years. For easy understanding, an example of equating MSPAP 1995 and 1996 scale scores is illustrated below (for details, refer to Maryland State Department of Education, 1996) .

The first step, called "Adjusting Test Form Effect" (refer to Figure 1), is to equate the three test forms using the linear equipercentile equating procedure under the assumption that the abilities of the three groups taking the three test forms are very similar. The second step, called

"Adjusting Rater Year Effect" (again refer to Figure 1), was taken to adjust for systematic effects in rater leniency or strictness in both year cohorts. In this equating, about 1,500 Answer Books per grade from the 1995 MSPAP administration were re-scored by some of the 1996 raters who were trained to re-score students' 1995 responses using the Scoring Guides developed for the 1995 MSPAP. Estimation of the rater effects was analyzed separately by content for each grade. The first set of scale scores ($95SS_{95}$) was based upon the ratings that the students had received from 1995 raters. The second set of scale scores ($95SS_{96}$) was based on the ratings that these students received from the 1996 raters. Both sets were expressed on the metric used for 1995 scale scores. Linear equipercentile equating procedures were used to estimate the transformation coefficients for rater year effect, which were used to transform the metric of $95SS_{96}$ into that of $95SS_{95}$.

The third step, called "Adjusting Yearly Test Version Effect" (see Figure 1), adjusted for systematic effects in test difficulty, which can be different in two year cohorts. This step was to identify a group of students in each grade who took the 1996 MSPAP and were equivalent to the 1996 group of students administered the 1995 MSPAP test. Linear equipercentile equating procedures were used to estimate the transformation coefficients for yearly test version effect, which were used for aligning the metric of MSPAP 1996 with the metric of MSPAP 1995. Finally, the MSPAP 1996 scale scores were transformed to the metric of the MSPAP 1995 scale score, using the transformation coefficients of test form effect, rater year effect as well as yearly test version effect.

While many statistical assumptions made for scaling (e.g. unidimensionality) and for test equating are unlikely to be exactly true in practice, especially for a performance-based program. Besides that, this type of assessment may encounter other practical problems, pointed out previously, even though efforts were made to avoid them.

[Insert Figure1 here]

II. Overview of Statistical Procedures

A. Association between the Performance-based and Multiple-choice Assessments

The degree of association between a performance-based and multiple-choice assessments is often used as a means to evaluate a new performance-based assessment program. However, test practitioners have faced a dilemma in interpreting the results generated from this type of statistical analysis. Do we expect to obtain high correlation between the two measures? High correlation might be a good indicator of validity. For instance, Yen (1998) investigated how CTBS5(Comprehensive Tests of Basic Skills) scores from the previous grade related to MSPAP proficiency. For Grade 2 students who took CTBS/5 reading and mathematics and who were rated at the level of "Proficient", 65 and 64 percent of these students one-year later were rated proficient on MSPAP reading and mathematics, respectively. However, when the performance-based assessment is strongly associated with what the multiple-choice assessment intends to measure, a question of the need for the time-consuming method of the performance-based assessment to assess student achievement can be raised.

On the other hand, do we expect to obtain a result with low correlation between two measures? Low correlation might be an indicator of the unique characteristics of the performance-based assessment as compared to the multiple-choice assessment. However, low correlation would be cause for concern, because the validity of the performance-based model would be called into question.

A more sophisticated approach to investigate the association between two measures is known as multitrait-multimethod (MTMM). It was developed by Campbell and Fiske (1959). This model includes four types of correlation in the following order of their results from largest to smallest (Nunnally & Bernstein, 1994):

(1). The correlation (reliability) between the same trait scores measured by similar methods.

(2). The correlation (validity) between the same trait scores measured by different methods.

(3). The correlation between two different trait scores measured by similar methods.

(4). The correlation between two different trait scores measured by different methods.

Schatz (1998) applied the MTMM approach to examine the reliability-validity coefficients for reading and mathematics achievement scores. Each content was assessed by two multiple-choice measures, CTBS/4 and a CRT(Criterion Referenced Test) and by one performance-based assessment, MSPAP. The expected order of correlation coefficients was found for the content area of Mathematics at three grade levels, Grade 3, 5 and 8. The validity coefficients for the content area of Reading did not fit the expected pattern at any of the three grade levels. Was this problem caused by the performance-based assessment or by the multiple-choice assessment? The answer to this question based on the analysis of MTMM correlation was unclear. In addition to that, visual inspection for assessment of construct validity data in a correlation matrix can be problematic because of measurement and sampling errors.

Using the degree of association between two different types of assessment models to evaluate the construct validity of the performance-based assessment, researchers generally encounter problems in reaching a conclusion. The structural equation modeling (SEM) (for literature review, see Schmitt & Stults,1986) may relieve part of the above problem. It is capable of further partitioning the variance of each content measure into three components: specific trait; assessment method; and random error. The comparisons among the magnitudes of the three components for each measure is another criteria to evaluate the construct validity of the performance-based or multiple-choice assessment. More technical details will be illustrated in the section of Methodology.

B. Longitudinal Association between two Performance-based Measures

An alternative to evaluate a performance-based assessment program is the longitudinal association technique between two performance-based measures; for instance, test scores for students who had multiple-subject scores on two performance-based assessments when they are in a current grade and in a previous-year grade. An intercorrelation analysis is performed. One might expect that the correlation between two performance-based measures of the same content area should be higher than its correlations with different content areas when the measure errors are appropriately taken into control. This type of analysis does not depend on different types of measures, so that the correlation obtained from this analysis is much easier to interpret than that from the association between two different-type measures..

III. Methodology

A. Longitudinal Associations of Grade 3 MSPAP with Grade 5 MSPAP

1. Data Description and Sample Size

Test scores for students who had six content area scores on both MSPAP measures when they were in third grade in 1994 and in fifth grade in 1996 were collected from the Prince George's County school district. Approximately 5,500 students' samples were available.

2. Data Analysis and Evaluation

The analysis of the intercorrelations among students' performance on the two time-period measures in six content areas was performed. The sampling error is minor, due to the relatively large sample size. However, the measurement errors (unreliability) of two measures, particularly in the performance-based measure, can not be avoided and will cause correlation attenuation (Lord, 1980).

A correction for attenuation can be obtained by computing the true-score (without measurement error) relationship between two tests. Technically, creating factors with only a single measured indicator variable is a tool to approximate the true-score correlation when structural equation modeling is applied. Consider the diagram in Figure 2. TRD96 and TWR96 represent the constructs underlying observed variables of MSPAP reading in 1996 (RD96) and MSPAP writing in 1996 (WR96), respectively. The corresponding error variances of the standardized-scale variable RD96 and WR96 can be approximated by 1-Reliability Coefficient. These values of error variances were fixed while estimating the correlation between TRD96 and TRD94. The internal reliability coefficient of Cronbach's alpha was available from the MSPAP technical report and used for approximating the error variance. Similar principles are applied to compute any pair of true-score correlation of any two tests. In essence, the true-score correlation of two measures depend on trustworthy reliability information. For the rest of the figures presented in this study, the rectangles and circles denote the observed variables and latent factors, respectively. The labels of RD, WR, LS, MS, SS and SC stand for the MSPAP reading, writing, language usage, mathematics, social study, and science, respectively. The numbers 96 and 94 denote the year. The symbols "E" and "D" represent the error term for the observed variable and residual term for the latent variable, respectively. The SEM computer program, EQS, (Bentler, 1995) was used to estimate the SEM parameters of interest.

Another alternative to minimize the effect of measurement error on estimating the intercorrelations between two measures is to use the school-based scores (school mean) instead of individual students' scores that were unreliable measures according to the MSPAP test construction as illustrated in the MSPAP technical report. This school-based correlation analysis is particularly meaningful for MSPAP.

Based on the above longitudinal association analyses, the following hypothesis was examined: the adjusted or group-mean correlation between two similar measures of the same content area should be higher than its correlations with different content areas (Nunnally & Bernstein, 1994).

[Insert Figure 2 here]

An exploratory factor analysis was explored, for instance, for the Grade 5 MSPAP data in 1996. In addition to that, structural equation modeling was conducted to attempt to partition the variance of each content area measure of MSPAP into the components: specific trait, measurement method and error term. Several specific SEM models are illustrated below. Model comparisons were performed to explore which model was better in terms of data-model fit. It is important to note that our model comparisons were by no means exhaustive. Other models may be of interest.

Model L1: Six Correlated Latent Traits

A model for the unadjusted intercorrelations in Table 1 is represented by the path diagram shown in Figure 3. Six latent variables representing the true scores on the six traits are postulated. For instance, the latent trait of READING is supposed to be measured by RD96 and RD94. In addition, these six latent traits are intercorrelated. Each observed measurement is assumed to be determined by a trait and an error term. The variance of the error term of the standardized-scale variable is constrained by 1- Reliability Coefficient, where the reliability coefficient is obtained as Cronbach's alpha value. The assumption behind this model is that the six intercorrelated latent traits and their corresponding measurement errors are capable of explaining the intercorrelation matrix being analyzed.

[Insert Figure 3 here]

Model L2: A Second-order Trait Model

Another model for the intercorrelations described previously is represented by the path diagram shown in Figure 4. Since the magnitudes of correlation among the six lower-order factors (latent traits) specified in Model L1 were relatively high, a higher-order factor (Labeling Second-Order F) rather than the correlation of these six traits among themselves was hypothesized to account for this correlation matrix. The variance of the error term was constrained by the method described previously. The similarity between this model and Model L1 is that only traits and error terms were specified in the model. In contrast, the models described below will include the Method Effects into the model. We tried to incorporate the method effects into Model L1. Unfortunately, the problem of linear dependence on some parameter estimates (refer to Bentler, 1995) was encountered. Accordingly, the model of L2 serves as the base line against which an alternative model, Model L3 presented below, is compared.

[Insert Figure 4 here]

Model L3: A Second-order Trait and Method Effects

Model L3 (see Figure 5) was formed by adding the Method effects into model M2. Model M2 is nested within Model L3. It is hypothesized that the six content measures from 1996 data reflect 1996 Method Effect (PAM96) and the six content measures from 1994 data reflect 1994 Method Effect (PAM94). Model comparison between this model and Model L2 was conducted to explore whether the Method Effects can significantly improve in fitting the data.

[Insert Figure 5 here]

Model L4: Modified Model L3 by freeing Several Error Variances

In order to improve the model-data fit several variances of error terms and a covariance of residual for the second-order factor analysis were set free to be estimated. They are specified in Table 4.

B. Multitrait-multimethod Associations of MSPAP, CTBS and OLSAT

1. Data Description and Sample Size

Students in third grade in 1996 had six content area scores of MSPAP, three content area scores (Reading Vocabulary Scale, RVS, Reading Comprehension Scale, RCS and Math Application Scale, MAS) of the CTBS and the Otis Lennon School Abilities (OLSAT). The CTBS and OLSAT are multiple-choice format instruments. The sample size is about 7,000.

2. Data Analysis and Evaluation

An intercorrelation analysis was conducted for the six content area scores of MSPAP, three content area scores of CTBS, and the OLSAT score. Similar correlation analyses were conducted using school-based mean statistics. Regarding the intercorrelation matrix in Table 5 (MSPAPRD, MSPAPMS, CTBSRVS, CTBSRCS and CTBSMAS), structural equation modeling was conducted. Four specific SEM models are illustrated below. Hypothesis test and fit indices are used to evaluate whether modes are attainable. Besides that, a test in difference chi-square values between two nested models is used to evaluate which model is capable of capturing the data. Finally, decomposing the variance of the reading or mathematics measures into the components: specific trait, measurement method effect, and error term, can be used to evaluate whether the assessment method effects attenuate the trait effect.

Model M1: Correlated Latent Traits and Correlated Method Effects

A base line model for the intercorrelation matrix is represented by the path diagram shown in Figure 6. Two correlated trait factors and two correlated method-effect factors are hypothesized to underline the correlation matrix. Specifically, it is hypothesized that latent trait of READING is measured by MSPAPRD (MSPAP reading), CTBSRVS (CTBS reading vocabulary) and CTBSRCS (CTBS reading comprehension). MSPAPMS (MSPAP mathematics) and CTBSMAS (CTBS mathematics application) are hypothesized to be indicators of another latent variable of MATH. It is hypothesized that MSPAPRD and MSPAPMS reflect Method of Performance-based assessment (called MSPAP) and CTBSRVS, CTBSRCS and CTBSMAS reflect Method of Multiple-choice assessment (called CTBS). This model serves as the base line against which an alternative model presented below is compared. It is typically the least restrictive model. The variances of the error terms for MSPAPRD, MSPAPMS and

CTBSMAS were constrained by the method described previously. The error term variance for the CTBSMAS was unavailable and was approximated (set to .20) in order to gain 1 degree of freedom. The variances of the error terms for CTBSRVS and CTBSRCS were free to be estimated since no reliability information was available for these two measures.

[Insert Figure 6 here]

Model M2: No Traits and Correlated Method Effects

Model M2 is nested within Model M1. No trait factors were specified in the model.

Model M3: Perfectly Correlated Traits and Correlated Method Effects

Model M3 was formed by fixing the correlation between two trait factors to 1.0 in Model M1.

Model M4: Correlated Traits and Perfectly Correlated Method Effects

Model M4 was formed by fixing the correlation between two method factors to 1.0 in Model M1.

Using Widaman's (1985) paradigm, the evidence of convergent validity can be tested by comparing a model in which traits are specified (Model M1) with one in which they are not (Model M2). A test of difference in chi-square values between the two models was conducted. A more specific assessment of the convergent validity can be ascertained by examining the variance components on each measure due to trait, method and error. Further scrutiny of the variance components might detect the likelihood for method effects to attenuate the trait effects.

In testing for evidence of discriminant validity among traits, a comparison between a model in which traits correlated freely (Model M1) with one in which they are perfectly correlated (Model M3) was made. A test of the difference in chi-square values between two models was conducted to evaluate the discriminant validity of traits.

The same logic, as noted earlier, was used to evaluate the evidence of discriminant validity among methods. A model in which method factors were freely correlated (Model M1) was compared with one in which they are perfectly correlated (Model M4). A test of the difference in chi-square values between two models was conducted to evaluate the evidence of discriminant validity of the method factor. Finally, we remind readers that our model comparisons were by no means exhaustive. Other alternative models may be of interest.

IV. Results and Discussions

A. Analyses of Longitudinal Associations of Grade 3 MSPAP with Grade 5 MSPAP

1. Three Types of Intercorrelation Matrix

The results of correlation analysis for students who had six content area scores on both MSPAP measures when they were in third grade in 1994 and in fifth grade in 1996 are presented in Table 1. As illustrated in the section on methodology, three types of correlation analysis were performed. The unadjusted correlation (labeled as UnAdj) is presented in the first row in each cell. The correction correlation for attenuation (labeled as Adj) is presented in the second row in each cell. The correlation calculated from school-mean (labeled as Group) is presented in the third row in each cell.

[Insert Table 1 here]

The values underlined represent the reliability coefficients which reflect the underlying-trait true correlations between the two same content measures across two years. Similarly, the values shown in bold-font represent the reliability coefficients based on school mean. The values in off-diagonal within the thick black borders represent the correlations of two measures of different content areas between MSPAP 1994 and 1996. One might expect that the correlation between two measures of the same content area should be higher than its correlations with different content areas. Unfortunately, this was not the case for all the content areas. For instance, the adjusted correlation between Read96 and Read94 was 0.620, which was smaller than the correlations of Read96 with Social Study94 (.650), Social Study96 (.657) and Science96 (.667). The hypothesis made in this study is not well held in the test data examined. Two questions are raised according to these results. One question is: Can this result be generalized to the test data of other school districts or the whole state? This question can be appropriately examined by analyzing the longitudinal data collected from the whole state school district. Another question is : Can this assumption be retained when the multiple-choice assessment program (for instance CTBS multiple-subject assessments) is applied? A future study of the longitudinal associations of the multiple-choice assessments should be conducted to serve as a base for comparisons with this study. Practically, if the answer to the latter question is "NO", one might wonder whether the hypothesis made in this study is unpractical for the multiple-subject assessment program. Meanwhile, the fact that the MSPAP test data being analyzed in this study violated this assumption becomes less serious than we originally thought. However, if the answer is "YES", the search for the reasons, for instance, scaling or test equating issues on MSPAP, will become critical.

2. Exploratory Factor Analysis

Further factor analysis on the correlation matrix of the set of six content area scores, for instance, Grade 5 MSPAP data in 1996, was conducted. It turned out that approximately 72 percent of the variance-covariance of these six content area scores was accounted by one latent trait. One possible reason for this finding is that the factor of the MSPAP test tasks being integrated both within a content area and across content areas may capture most of the common variance among the six content scale scores.

Another possible reason is that this common variance may account for a general ability (Cronbach, 1970). Accordingly, the estimate of the proportion of the unique variance for each content-area measure will be a valuable index to reflect the efficacy of a specific content-area measure. The proportion of unique variance for each content area can be estimated by subtracting the proportion of error variance from corresponding proportion of unexplained variance (or unique and error variances) that equals one minus the value of community. The estimated proportion of error variance for each content test can be approximated by one minus the corresponding coefficient Alpha (from MSPAP 1996 technical report). Finally, the proportions of unique variance for Reading (0.05), Writing (0.09), Social Studies (0.06) and Science (0.09) are very low (see Table 2).

[Insert Table 2 here]

The finding from the exploratory factor analysis is not consistent with results from literature on factor analysis studies, in which the verbal oriented ability tests such as Reading, Writing and Language usage and math oriented ability tests such as Math and Science are usually separately factored by two different underlying traits. Further analyses using structural equation modeling (Bollen, 1989) were conducted and will be presented below.

3. Hypothesis tests for Models L1 to L4 and Model Comparisons

The hypothesis tests for Models L1 to L4 are presented in Table 3. Models L1 and L2 poorly fit the data in terms of fit indices. Both hypotheses (made in Model L1: the six correlated traits themselves are capable of accounting for the correlation matrix being analyzed, and made in Model L2: a higher-order trait can capture the intercorrelations among the six traits) are not attainable. However, if the method effects were added into the model L2, a significant increase in data-model fit was found in Model L3 (see Table 3). Hu and Bentler (1999) recommended joint criteria to retain a model, such as (CFI>=.96 and SRMR <=.10) or (RMSEA<=.06 and SRMR <=.10). Model L4, freeing some error term variances in Model L3, is retained because it meets any of these two joint criteria.

[Insert Table 3 here]

According to Model L4, the variance components due to Trait, Method Effect, and Error for the first-order Factor Analysis is presented in Table 4. The method effects play a substantive role in accounting for the variance of the six latent traits. Similar variance components due to the second-order factor and residuals for the second order Factor Analysis is presented in Table 4. The latent traits of Reading, Math, Social Study and Science had very high loadings on the higher-order factor.

[Insert Table 4 here]

B. Multitrait-multimethod Associations of MSPAP, CTBS and OLSAT

1. Two Types of Intercorrelations

The intercorrelation analyses among MSPAP, CTBS and OLSAT measures for 1996 test data were carried out and their results are presented in Table 5. As noted in the section of methodology, two types of correlation analysis were performed. The unadjusted correlation (labeled as UnAdj) is presented in the first row in each cell. The correlation calculated from school-mean (labeled as Group) is presented in the second row in each cell. For the reading measure of MSPAP, it is almost equally correlated with the reading vocabulary, reading comprehension and math application of the CTBS test and the general ability measure of OLSAT. A similar finding was found for the mathematics measure of MSPAP.

[Insert Table 5 here]

2. Hypothesis Tests for Models M1 to M4 and Model Comparisons

The chi-square value for the hypothesized model M1 (Correlated Traits and Correlated Methods) is .21 (see Table 6). The corresponding type I error is .645, indicating that this model fit data. In addition, this model is attained according to the two joint criteria (Hu & Bentler, 1999).

The chi-square value and the goodness-of-fit statistics for the Model M2 (No Traits and Correlated Methods) are presented in Table 6. As indicated by the chi-square and the fit indices, the goodness of fit for Model M2 was poor.

[Insert Table 6 here]

The evidence of convergent validity was tested by comparing a model in which traits are specified (Model M1) with one in which they are not (Model M2) using Widaman's (1985) paradigm. A significant difference in chi-square values between the two models supports evidence of convergent validity as happened here (see Table 6). A more specific assessment of the convergent validity can be ascertained by examining the variance components on each measure due to trait, method and error (see Table 7). Further scrutiny of the variance components reveals the likelihood for method effects to attenuate the trait effects. For instance, the Method effect might play a substantive role in accounting for the variance of MSPAP reading. This result might help us interpret the finding from Schatz's (1998) study, that the validity coefficients for the content area of Reading did not fit in the expected order in correlation coefficients, illustrated in the section on literature review. The trait effect of the CTBS mathematics application was also attenuated by the multiple-choice assessment method. The results from the variance component analysis seem to imply that either the performance-based assessment or the multiple-choice assessment can attenuate the trait effects.

[Insert Table 7 here]

The chi-square value and the goodness-of-fit statistics for Model M3 (Perfectly Correlated Traits and Correlated Methods) are presented in Table 6. We see that the fit of this model is fairly good, albeit slightly less well fitting than for Model M1. In testing for evidence of discriminant validity among traits, a significant difference in chi-square values between Model 1 and Model 3 was found (see Table 6) to support evidence of discriminant validity of traits.

The chi-square value and the goodness-of-fit statistics for the Model M4 (Correlated Traits and Perfectly Correlated Methods) are presented in Table 6. The fit of this model is almost as good as Model M3, albeit slightly less well fitting than for Model M1. In testing for evidence of discriminant validity among methods, we applied the same logic as noted earlier. A significant difference in chi-square values between these two models of M1 and M4 was found (see Table 6) to support evidence of discriminant validity of the method factor.

V. Summary and Conclusion

The primary concern of this study was to examine the construct validity of MSPAP by means of analyzing the performance-based test data set in one school district. Based on the analyses of the longitudinal associations of Grade 5 MSPAP data in 1996 with Grade 3 MSPAP data in 1994, the following hypothesis was examined: the unattenuated correlation or the group-mean correlation between two similar measures of the same content area is higher than its correlations with different content areas. This hypothesis was not attained. Although this finding might threaten the construct validity of MSPAP and bring the broad question of whether the content-area scores obtained on MSPAP reflect the efficacy of the instructional programs delivered in schools, school districts, and the state, we would be very prudent not to prejudge this issue because of two questions associated with this finding. The questions are: (1). Can this result be generalized to the test data of other school districts or the whole state?; and (2). Can this assumption be retained when the multiple-choice assessment program (for instance CTBS multiple-subject assessments) is applied? These two questions need clarification at some future time.

In addition, the results analyzed by structural equation modeling (SEM) to this longitudinal correlation matrix reveal that the SEM model specified by the MSPAP six latent traits was unable to capture the underlying information of this data. Extra factors, such as, a general ability and an assessment method effect, may need to be considered for better fitting the data. This result seems to imply that what we have observed in MSPAP data is more general measures of student ability than their performance in any given content area.

The results from structural equation modeling to the multitrait-multimethod correlation data suggest that the trait effect of MSPAP reading may be attenuated by the method effect of the performance-based assessment . Similarly, the trait effect of CTBS mathematics application may also be attenuated by the method effect of the multiple-choice assessment. These phenomena of trait effects attenuation by assessment methods can happen in either performance-based or multiple-choice assessment. The issue of whether these findings can be generalized to other MSPAP data is worthwhile to investigate at some future time.

The primary rationale for moving away from the multiple-choice assessment to the performance-based assessment comes from a strong belief that student "assessment needs to mirror instruction and high-quality learning activities" (p 53, Linn, 1995). This movement is motivated primarily by instructional rather than psychometric considerations. The current psychometric techniques such as test equating and scaling that have predominately been used for the multiple-choice assessment for a long time may not completely suitable to this new assessment movement. New psychometric techniques such as multidimensional scaling (Ackerman, 1994; Reckase, 1997) and equating (Li & Lissitz, in press) will serve as an important tool to quantify this type of assessment when at some future time these new statistical techniques resolve the problems they are facing.

References

Ackerman, T. A. (1994). Using multidimensional item response theory to understand what items and tests are measuring. Applied Measurement in Education, 4, 255-278.

Bentler P. M. (1995). EQS Structural equations program manual. Encino, CA: Multivariate Software, Inc.

Bollen, K. A. (1989). Structural equations with latent variables. New York: A Wiley-Interscience Publication.

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

Cronbach, L. J. (1970). Essentials of psychological testing. New York: Harper & Row Publishers, Inc.

Green, B. F. (1995). Comparability of scores from performance assessments. Educational Measurement: Issues and Practice, Winter, 13-15.

Hu, L.& Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling: A multidisciplinary Journal, 6, 1-55.

Li, Y. H. & Lissitz, R. W. (in press). An evaluation of multidimensional IRT equating methods by assessing the accuracy of transforming parameters onto a target test metric. Applied Psychological Measurement.

Linn, R. L. (1995). High-stakes uses of performance-based assessments: Rationale, examples, and problems of comparability. In T. Oakland & R. K. Hambleton (Ed.), International perspectives on academic assessment (pp. 49-73).Norwell, MA. Kluwer Adademic Publishers.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. New Jersey: Lawrence Erlbaum Associates, Inc.

Maryland State Department of Education. (1996). Technical report: 1996 Maryland School Performance Assessment Program. Baltimore: Author.

Nunnally, J. C. & Bernstein, I. H. (1994). Psychometric theory. New York: McGraw-Hill, Inc.

Reckase, M. D. (1997). The past and future of multidimensional item response theory. Applied Psychological Measurement. 21, 25-36.

Schatz, C. J. (1998, November). Convergent-discriminant validity evidence for the MSPAP Reading and Math scores. Paper presented at the annual meeting of the Maryland Assessment Group, Ocean City, MD.

Schmitt, N. & Stults, D. M. (1986). Methodology review: Analysis of multitrait-multimethod matrices. Applied Psychological Measurement, 10, 1-22.

Widaman, K. F. (1985). Hierachically tested covariance structure models for mulrait-multimethod data. Applied Psychological Measurement, 9, 1-26.

Yen, W. M. & Ferrara, S. (1997). The Maryland school performance assessment program: Performance assessmetn with psychometric quality suitable for high stake usage. Educational and Psychological Measurement, 57, 60-84.

Yen, W. M. & Ferrara, S. (1997). The technical quality of performance assessments: Standard errors of percents of pupils reading standards. Educational Measurement: Issues and Practice, Fall, 5-15.

Yen, W. M. & Julian, M. W. (1998, November). How CTBS/5 scores from the previous grade relate to MSPAP proficiency. Paper presented at the annual meeting of the Maryland Assessment Group, Ocean City, MD.

Figure Headings

Figure 1. The Equating Design Used for Equating MSPAP 1995 and 1996 Scale Scores

Figure 2: A SEM model for Computing the Intercorrelations among the Twelve True Scores of MSPAP

Figure 3: A SEM Model: Six Correlated Latent Traits for a MSPAP Longitudinal Associations Data

Figure 4: A SEM Model: A Second-order Factor for a MSPAP Longitudinal Associations Data

Figure 5: A SEM Model: A Second-order Factor and Method Effects for a MSPAP Longitudinal Associations Data

Figure 6: A Hypothesized Multitrait-multimethod Model for the MSPAP-CTBS Correlation Data

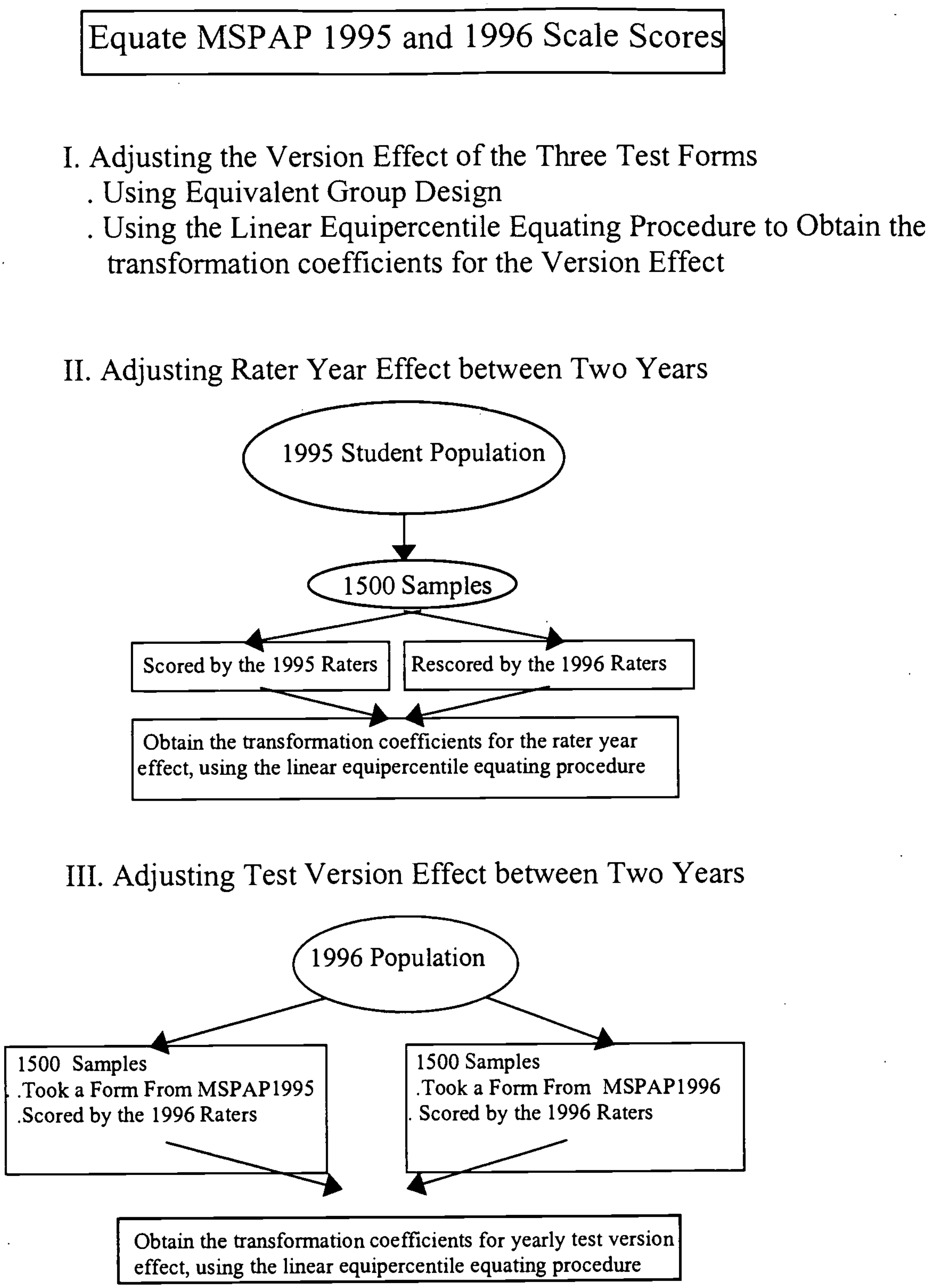
Equate MSPAP 1995 and 1996 Scale Scores
I. Adjusting the Version Effect of the Three Test Forms
. Using Equivalent Group Design
. Using the Linear Equipercentile Equating Procedure to Obtain the transformation coefficients for the Version Effect
II. Adjusting Rater Year Effect between Two Years
1995 Student Population
1500 Samples
Scored by the 1995 Raters
Rescored by the 1996 Raters
Obtain the transformation coefficients for the rater year effect, using the linear equipercentile equating procedure
III. Adjusting Test Version Effect between Two Years
1996 Population
1500 Samples
.Took a Form From MSPAP1995
.Scored by the 1996 Raters
1500 Samples
.Took a Form From MSPAP1996
. Scored by the 1996 Raters
Obtain the transformation coefficients for yearly test version effect, using the linear equipercentile equating procedure

Figure 1. The Equating Design Used for Equating MSPAP 1995 and 1996 Scale Scores

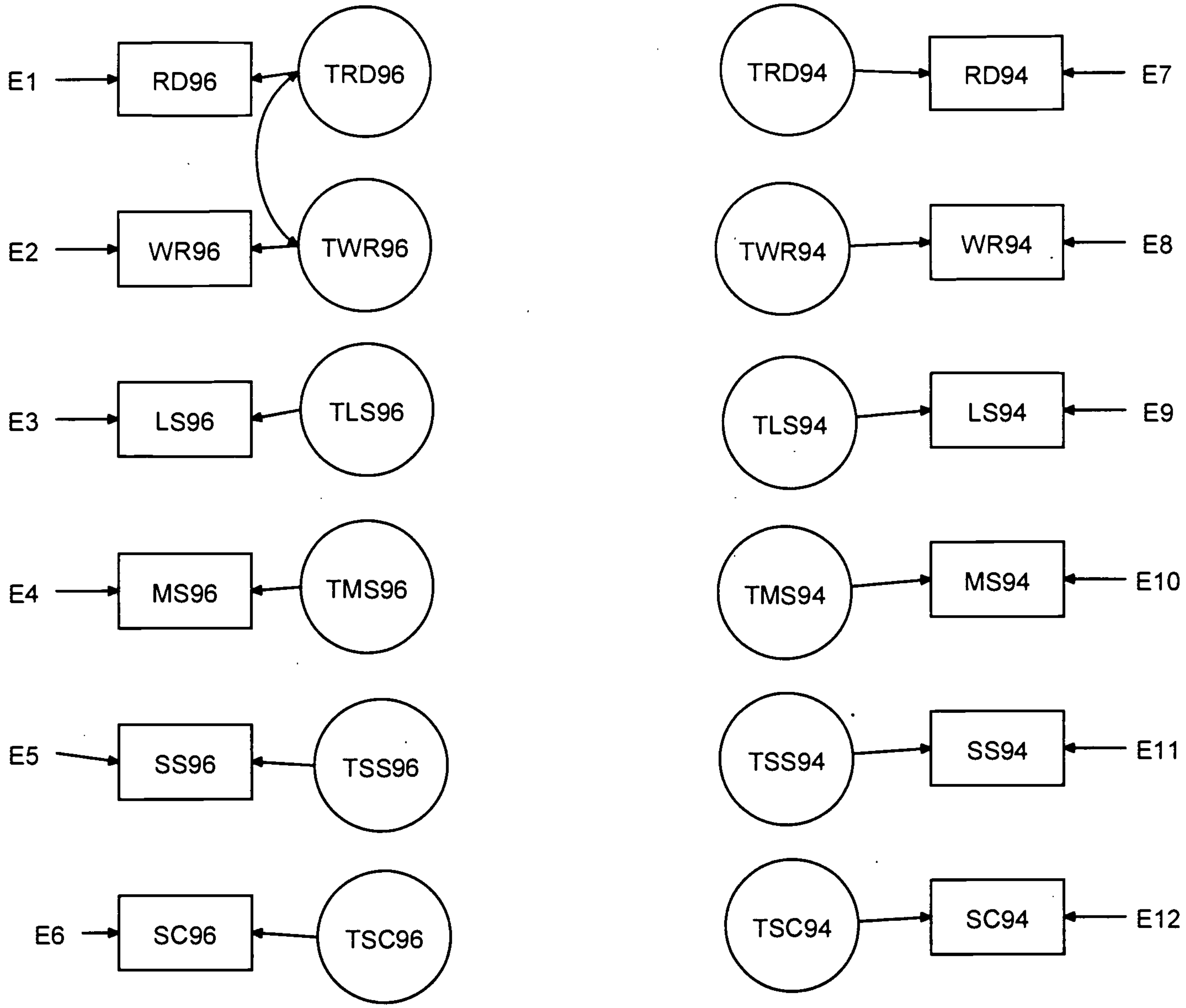

Figure 2: A SEM model for Computing the Intercorrelations among the Twelve True Scores of MSPAP

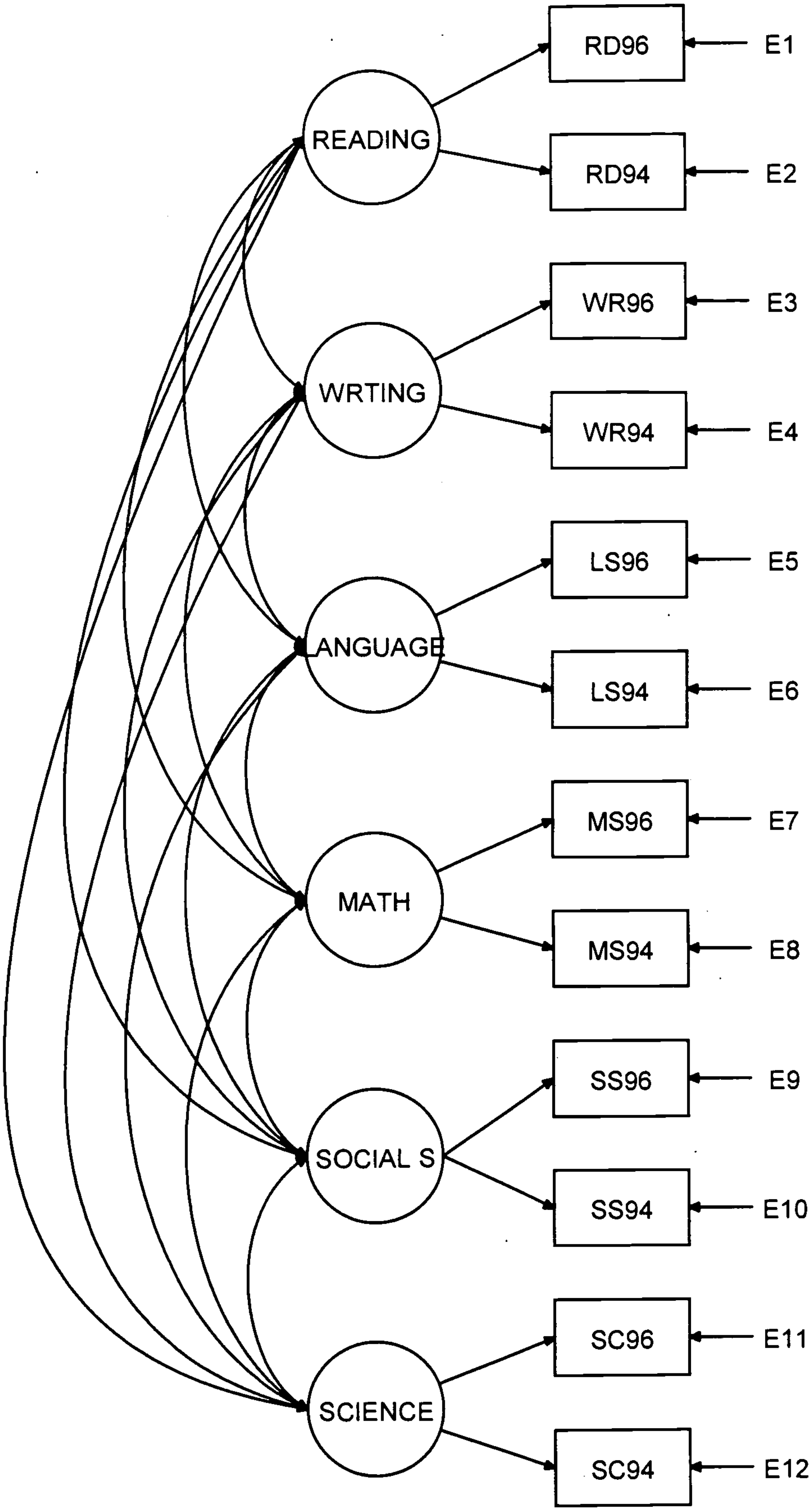

Figure 3: A SEM Model: Six Correlated Latent Traits for a MSPAP Longitudinal Associations Data

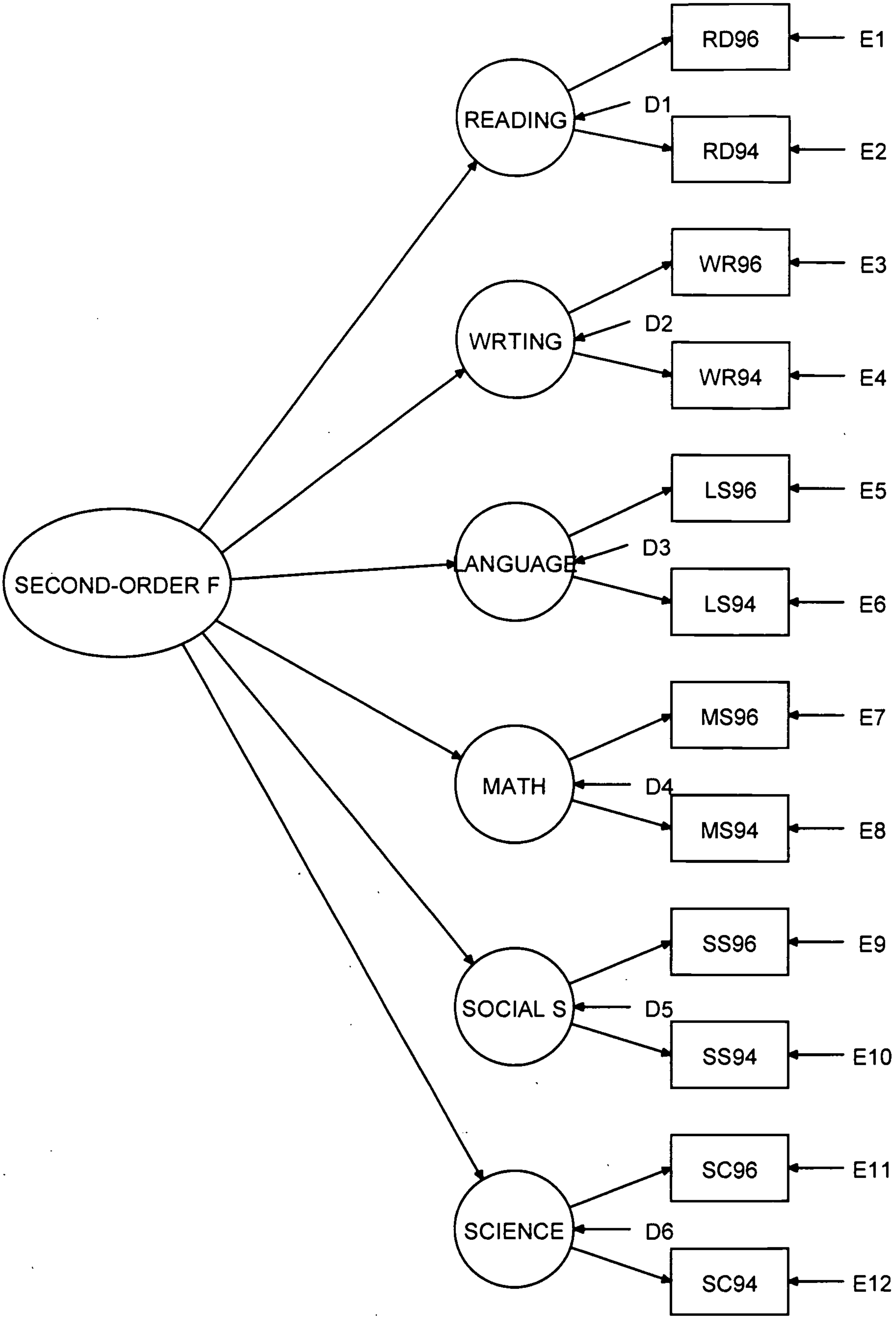

Figure 4: A SEM Model: A Second-order Factor for a MSPAP Longitudinal Associations Data

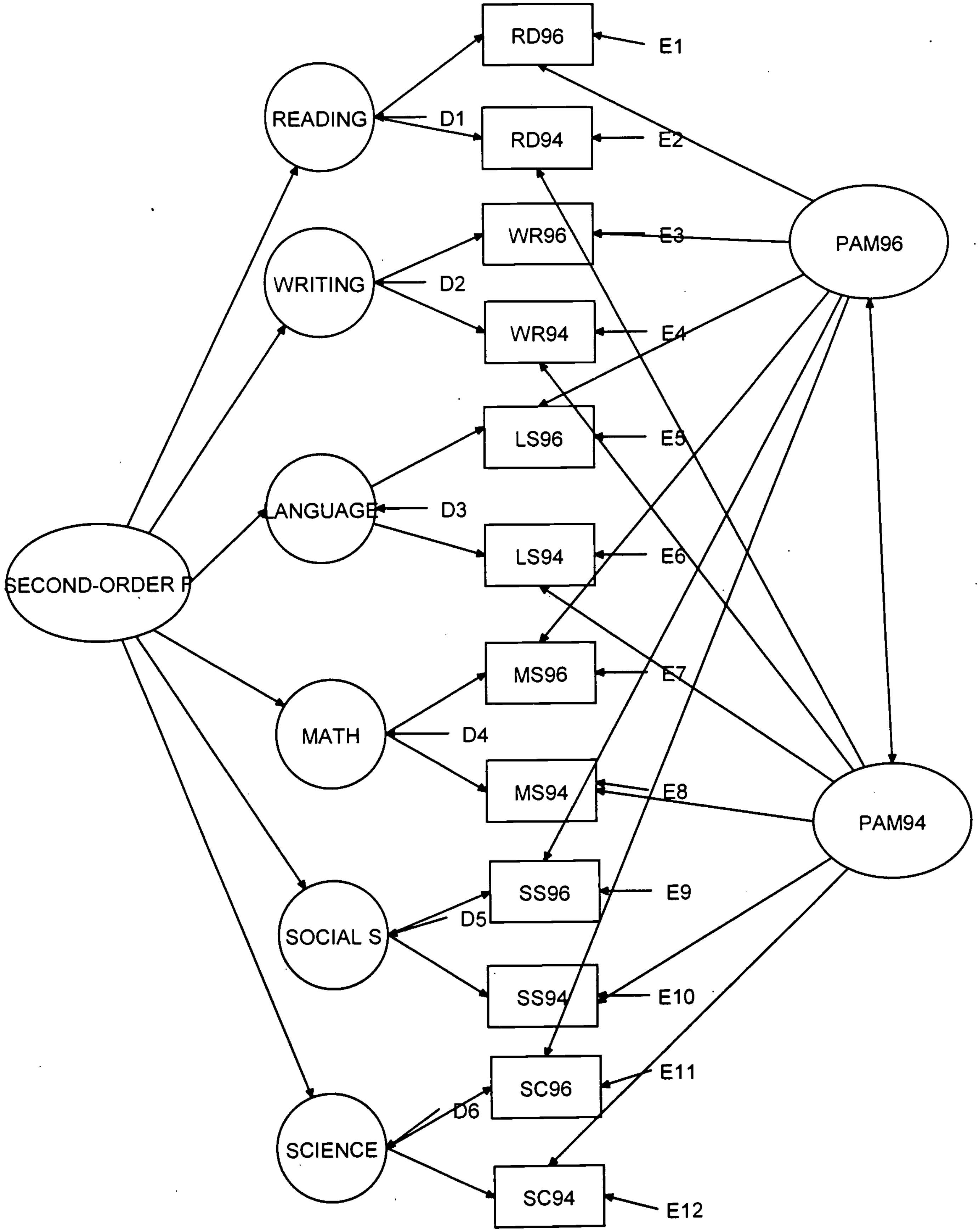

Figure 5: A SEM Model: A Second-order Factor and Method Effects for a MSPAP Longitudinal Associations Data

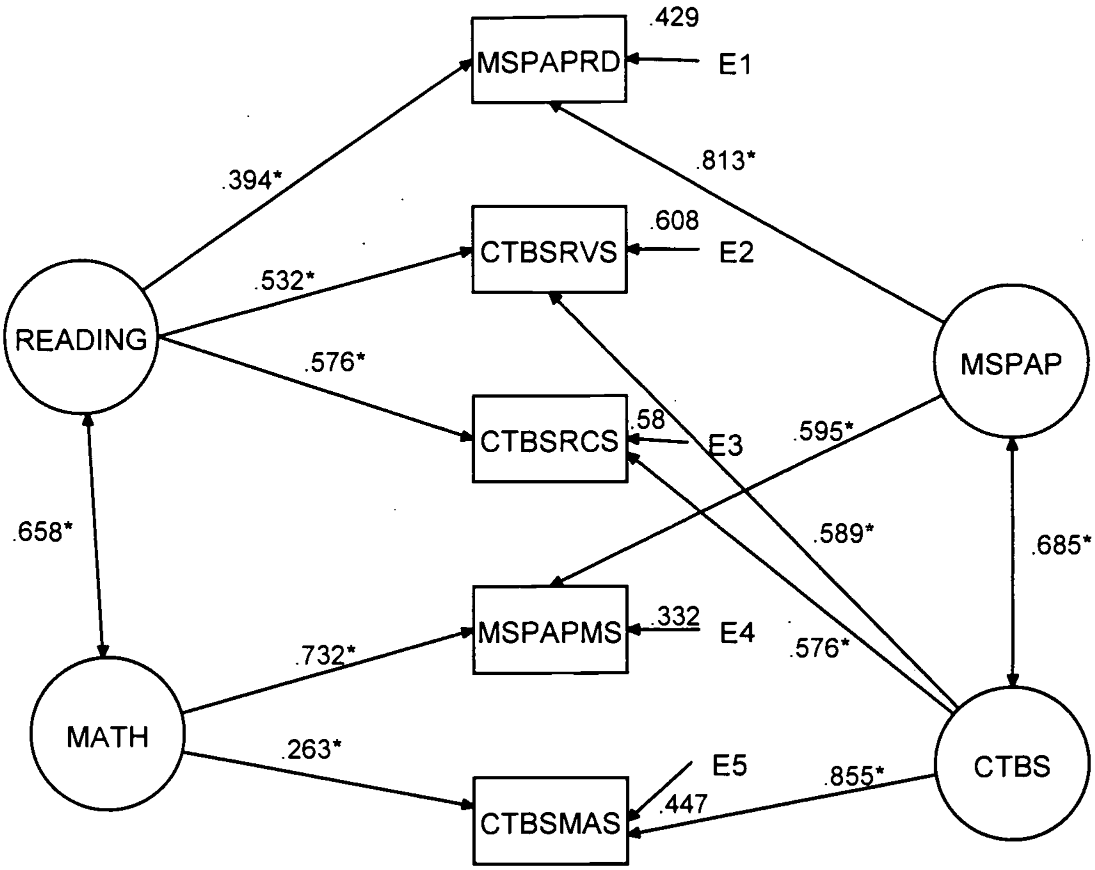

Figure 6: A Hypothesized Multitrait-multimethod Model for the MSPAP-CTBS Correlation Data

Table 1.The Intercorrelation Matrix of a Longitudinal Data, Grade 3 MSPAP 1994 with Grade 5 MSPAP 1996

Content	Type of Corre.	RD96	WR96	LS96	MS96	SS96	SC96	RD94	WR94	LS94	MS94	SS94	SC94
RD96		1.00											
WR96	UnAdj Adj Group	.506 .676 .842	1.00										
LS96	UnAdj Adj Group	.574 .673 .864	.717 .887 .902	1.00									
MS96	UnAdj Adj Group	.602 .727 .795	.534 .679 .821	.586 .655 .763	1.00								
SS96	UnAdj Adj Group	.669 .812 .851	.571 .731 .864	.615 .692 .847	.663 .767 .888	1.00							
SC96	UnAdj Adj Group	.748 .908 .900	.573 .733 .893	.630 .708 .862	.744 .860 .913	.725 .843 .931	1.00						
RD94	UnAdj Adj Group	.493 .620 **.731**	.444 .589 .719	.509 .593 .766	.502 .602 .639	.545 .657 .770	.553 .667 .722	1.00					
WR94	UnAdj Adj Group	.363 .561 .531	.387 .630 **.528**	.448 .641 .557	.398 .586 .467	.414 .613 .570	.432 .639 .535	.446 .685 .757	1.00				
LS94	UnAdj Adj Group	.456 .541 .625	.481 .602 .613	.592 .651 **.682**	.481 .544 .518	.500 .668 .635	.518 .589 .613	.542 .638 .827	.678 .981 .842	1.00			
MS94	UnAdj Adj Group	.467 .560 .656	.437 .553 .657	.504 .560 .701	.553 .632 **.614**	.529 .609 .695	.567 .651 .683	.584 .696 .852	.475 .695 .778	.550 .618 .828	1.00		
SS94	UnAdj Adj Group	.536 .650 .684	.492 .630 .694	.559 .628 .726	.554 .640 .600	.593 .689 **.721**	.608 .707 .699	.701 .845 .895	.517 .765 .800	.607 .690 .855	.676 .778 .900	1.00	
SC94	UnAdj Adj Group	.457 .571 .583	.426 .562 .598	.484 .561 .634	.524 .624 .595	.509 .610 .650	.541 .648 **.625**	.593 .736 .808	.460 .757 .767	.555 .650 .806	.667 .790 .877	.668 .801 .862	1.00

Table 2. Factor Loadings of Grade 5 MSPAP 1996 Data, Coefficient Alpha and the Proportion of Unique Variance for Each Content Area Test

	RDSS	WRSS	LSS	MSS	SSSS	SCSS	Explained Variance
Common Factor I Loading	.86	.79	.83	.84	.89	.87	72.30
Communality	.74	.62	.70	.71	.80	.77	
Unique & Error Variances	.26	.38	.30	.29	.20	.23	
Coefficient Alpha	.82	.71	.92	.89	.87	.83	
Error Variance	.21	.29	.08	.13	.14	.14	
Content Area Variance	.05	.09	.22	.16	.06	.09	

Table 3
Hypothesis Tests and Fit Indices for Models from L1 to L 4

Model	Chi-square	N	df	p	CFI	SRMR	RMSEA
L1 Six Correlated Latent Traits	35699.13	5500	51	.001	.210	.091	.357
L2 Second-order Factor	37299.20	5500	60	.001	.175	.097	.336
L3 Second-order Factor and Method Effects	7151.06	5500	47	.001	.843	.045	.166
L4 Modified Model L3 by freeing Some Error-term Variances or Covariances	740.99	5500	40	.001	.984	.021	.056

Model Comparison	Chi-square Difference	df	p	Better Fit Model
L3 vs L2	30148.14	13	.001	Model L3
L4 vs L3	6410.07	7	.001	Model L4

Table 4
Variance Components due to Trait, Method and Error for the first-order Factor Analysis and Variance Components due to the Second-order Trait and Residual for the second-order Factor Analysis

First –order Factor Analysis							Second –order Factor Analysis		
Content	Trait		Method		Error		First-order Trait	Second-order Trait	Residual
RD96	.23	Reading	.42	PAM96	.35		Reading	.96	.04
RD94	.45	Reading	.17	PAM94	.38				
WR96	.52	Writing	.19	PAM96	.30	f	Writing	.51	.49
WR94	.12	Writing	.41	PAM94	.47	f			
LS96	.71	Language	.21	PAM96	.08	f	Language	.46	.54
LS94	.17	Language	.72	PAM94	.10	f			
MS96	.29	Math	.38	PAM96	.33		Math	.89	.11
MS94	.47	Math	.19	PAM94	.34				
SS96	.29	Social S	.39	PAM96	.32		Social Study	.90	.10
SS94	.64	Social S	.21	PAM94	.14	f			
SC96	.30	Science	.56	PAM96	.14	f	Science	.98	.02
SC94	.41	Science	.20	PAM94	.39				

Note:

The correlation between PAM96 and PAM94 is .604

The symbol "f" represents the error term variance to be fixed as one minus the value of the Reliability Coefficient.

The covariance of the Residual for Writing and Residual for Language is free to be estimated.

Table 5.Multitrait-multimethod Correlation Matrix, among MSPAP, CTBS and OLAST

Content	Type of Corre.	RD96	WR96	LS96	MS96	SS96	SC96	RV96	RC96	MA96	OLRS96
RD96		1.00									
WR96	UnAdj	.570	1.00								
	Group	.871									
LS96	UnAdj	.625	.755	1.00							
	Group	.859	.900								
MS96	UnAdj	.674	.552	.593	1.00						
	Group	.922	.822	.794							
SS96	UnAdj	.771	.611	.678	.729	1.00					
	Group	.950	.880	.856	.927						
SC96	UnAdj	.744	.579	.629	.737	.752	1.00				
	Group	.942	.867	.842	.941	.941					
RV96	UnAdj	.537	.451	.552	.497	.562	.522	1.00			
	Group	**.589**	.561	.728	.523	.566	.561				
RC96	UnAdj	.549	.445	.541	.512	.563	.536	.646	1.00		
	Group	**.607**	.566	.708	.558	.572	.583	.810			
MA96	UnAdj	.544	.440	.527	.541	.558	.539	.596	.592	1.00	
	Group	.632	.602	.749	**.596**	.622	.611	.848	.872		
OLRS96	UnAdj	.535	.431	.503	.577	.548	.547	.537	.571	.643	1.00
	Group	.597	.552	.690	.564	.589	.594	.793	.818	.887	

Table 6
Hypothesis Tests and Fit Indices for Models from M1 to M4

Model	Chi-square	N	df	p	CFI	SRMR	RMSEA
M1 Correlated Traits & Correlated Methods	.21	7000	1	.645	1.000	.001	.000
M2 No Traits & Correlated Methods	3901.47	7000	2	.001	.760	.065	.282
M3 Perfectly Correlated Traits & Correlated Methods	569.92	7000	2	.001	.965	.032	.201
M4 Correlated Traits & Perfectly Correlated Method	569.30	7000	2	.001	.965	.032	.201
Model Comparison	Chi-square difference		df	p			
Test of Convergent Validity							
M1 vs M2	3901.26		6	.001			
Test of Discriminant Validity							
M1 vs M3 (traits)	569.71		1	.001			
M1 V2 M4 (methods)	569.09		1	.001			

Table 7

Variance Components due to Trait, Method and Error for Model M1

Content	Trait		Method	Error	
MSPAP Reading	Reading	.16	.66	.18	fixed
CTBS RVS	Reading	.28	.35	.37	free
CTBS RCS	Reading	.33	.33	.34	free
MSPAP Math	Math	.54	.35	.11	fixed
CTBS MSA	Math	.07	.73	.20	fixed

The Construct Validity of a Performance-based Assessment Program

Yuan H. Li, Valeria Ford, Leroy J. Tompkins

Prince George's County Public Schools, Maryland

Paper presented at the annual meeting of the American Educational Research Association, April, 19-23, Montreal, Canada

44

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC®

TM029846

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: The Construct Validity of a Performance-based Assessment Program	
Author(s): Yuan H Li, Valeria Ford, Leroy J. Tompkins	
Corporate Source:	Publication Date: 4/99

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

☒

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A

□

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

□

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Docum
If permission to re

Name: Yuan H. LI
Address: Prince George's County Public Schools
Room 205
Upper Marlboro, MD. 20772
Tel: 301-952-6764
Fax: 301-952-6228
Email: jeffli@pgcps.org

I hereby grant to the Educational Resou
as indicated above. Reproduction fro
contractors requires permission from the
to satisfy information needs of educatc

Sign here, please

Signature:	Printed Name/Position/Title:	
Organization/Address:	Telephone:	FAX:
	E-Mail Address:	Date: 8/14/99

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:

Address:

Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:

Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

THE UNIVERSITY OF MARYLAND
ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
1129 SHRIVER LAB, CAMPUS DRIVE
COLLEGE PARK, MD 20742-5701
Attn: Acquisitions

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
1100 West Street, 2nd Floor
Laurel, Maryland 20707-3598

Telephone: 301-497-4080
Toll Free: 800-799-3742
FAX: 301-953-0263
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com

PREVIOUS VERSIONS OF THIS FORM ARE OBSOLETE.