Multitrait-Multimethod Comparisons of Selected and Constructed Response Assessments of Language Achievement

Joseph J. Stevens

Patricia S. Clauser

University of New Mexico

Address Correspondence to Joe Stevens, 113 Simpson Hall, University of New Mexico, Albuquerque, NM 87131; (505)277-4203 (w), (505)292-1437 (h); e-mail: jstevens@unm.edu

Abstract

Two instruments using different assessment formats, selected-response and constructed-response, were examined to determine whether they measured the same constructs. The instruments were both state-mandated and were administered to elementary school students in a large urban school district. The first instrument consisted of the six, selected-response, language subtests of the Iowa Tests of Basic Skills. The second assessment was a writing portfolio that yields four analytically scored measures of children's writing on a single capstone piece from the portfolio. Each instrument purports to measure the same facets of language achievement. Scores for two samples of children on both instruments were analyzed with confirmatory factor analysis models that tested the effects of constructs and formats using a multitrait-multimethod (MTMM) framework. Results showed that a substantial proportion of subscore variance was associated with assessment format and unique variance, especially for the constructed-response format. The findings show that tests that may appear to be measuring the same constructs may in fact be measuring substantially different things when different assessment formats are used.

Multitrait-Multimethod Comparisons of Selected and Constructed Response Assessments of Language Ability

In recent years there has been increasing interest in the use of alternative item and assessment formats for many testing purposes. While large-scale testing was once well characterized by tests with exclusively multiple-choice and other selected-response formats, many tests and assessments in use today are “blended” instruments that include a mix of item types and formats or are formed of constructed-response, performance assessment, or other alternative formats (Bond, Braskamp, van der Ploeg, & Roeber, 1996). CCSSO (1999) reports that almost all states and jurisdictions now use instruments with a variety of item formats. Of 53 states and jurisdictions surveyed, selected-response items were used in 94%, extended constructed-response in 79%, short constructed-response in 51%, written assessment in 66%, and performance assessment in 38%.

While there is a great deal of interest in and support for the use of blended instruments and alternative formats, it is important to demonstrate that alternative formats measure well the traits they are intended to represent. If the intent is to measure the same traits or constructs, then there should be substantial convergent validity with traditional formats and relatively little variance of scores associated with the particular format used. Messick (1993) described this issue as one of “trait equivalence”. When convergent validity is present, construct-relevant variance should accumulate across formats or methods, thereby strengthening and generalizing construct validity of use and interpretation. When convergent validity is absent, however, construct validity is undermined through the intrusion of construct-irrelevant method and format effects.

On the other hand, the intent may not be to measure the same traits or constructs. Some have argued that one of the primary motivations for using alternative assessment methods and formats is that different cognitive processes are engaged or that different traits or constructs can be measured (e.g., Collins, 1990; Fredricksen, 1984; Gardner, 1992; Moss, 1994; Nickerson, 1989; Resnick & Resnick, 1992; Traub, 1993; Wolf, Bixby, Glenn, & Gardner, 1991). In this case, it seems reasonable to argue that there should be more moderate levels of convergent validity across methods and, in fact, some significant degree of discriminant validity should be present. It is still necessary to demonstrate for the related but distinct methods that there is construct validity of the separable uses and interpretations of the different instruments.

The research evidence is unclear, however, regarding the effects of item and response formats in alternative assessments. Some studies have indeed shown that different traits or factors are measured through the use of alternative formats (Ackerman & Smith, 1988; Bennett, et al, 1990; Bridgeman & Rock, 1993; Stevens & Clauser, 1995; Thissen, Wainer, & Wang, 1994; Ward, 1987; Werts, Breland, Grandy, & Rock, 1980). Other studies have found that the same traits or factors are measured regardless of assessment format or that little information is added through the use of alternative formats (e.g., Bennett, Rock, & Wang, 1991; Hancock, 1994; Lukhele, Thissen, & Wainer, 1994).

Recent studies have provided only limited evidence of how and whether alternative assessment methods may measure different traits, constructs, or may engage different cognitive processes than traditional methods (Traub, 1993). Several studies have found that, to some degree, different cognitive processes are elicited through the use of alternative formats. Using think-aloud protocols, Campbell (2000) found that constructed response formats elicited a greater degree of constructive thinking. Ward, Dupree, and Carlson (1987) and Martinez and Katz (1996) each reported format effects in the use of solution strategies. Lu and Suen (1995) found differences between assessment method and test-takers' cognitive styles. Katz, Bennett, & Berger (2000), however, failed to find differences in solution strategies across assessment formats; instead finding that format differences could be attributed to differences in reading comprehension.

Examination of the psychometric properties of alternative item and assessment formats has also raised concerns about alternative assessment methods (e.g., Hambleton, et al, 1995; Koretz, Stecher, Klein, & McCaffrey, 1994), including the adequacy of reliability and generalizability (Dunbar, Koretz, & Hoover, 1991; Shavelson, Baxter, & Gao, 1993) and convergent validity (Baxter, Shavelson, Goldman, & Pine, 1992; Burger & Burger, 1994; Stecher, et al, 2000).

Campbell (1960) and Messick (1993) discuss the use of the original Multitrait-Multimethod (MTMM) conceptualization of Campbell and Fiske (1959) for the testing of issues of trait and construct equivalence. Through the measurement of two or more traits using two or more assessment methods, the original MTMM method assesses convergent and discriminant validity through inspection and analysis of correlation matrices. Measures of the same trait should correlate highly, providing evidence of convergent validity. Measures of different traits should produce sufficiently lower coefficients to establish discriminant validity. Method effects are present when correlations within a method are larger than correlations across methods but within traits. Traub (1993) argues that evidence from "...multitrait-multimethod studies is necessary to the conclusion that different item formats assess somewhat different characteristics..." (p. 33).

Because of the difficulties in directly interpreting correlation coefficients and matrices, a number of statistical methods have been explored for the analysis of MTMM matrices; most commonly applied have been variations of factor analysis. Werts, Breland, Grandy, and Rock (1980) found the presence of an essay method effect using factor analytic methods. Ackerman and Smith (1988) found that essay responses contained reliable variance not associated with variability in multiple-choice scores measuring the same qualities of language performance. A number of previous studies, however, have employed either exploratory factor analysis or item response theory methods to evaluate format effects. Confirmatory Factor Analysis (CFA) models are a more powerful and direct means to establish the relative contributions of traits and methods in accounting for relationships among measures. Stevens and Clauser (1995) used CFA methods and found that relationships among language measures were best characterized by a model composed of correlated traits and correlated methods. Results showed that over 50% of the measured variance in subscores was related to the method of measurement (multiple-choice versus constructed-response) rather than the language traits being measured.

The purpose of the present study was to apply CFA methods using two samples to examine the concurrent and discriminant validity of a writing portfolio and the language subtests of a selected-response instrument that reference the same constructs. Our primary interest was in using the MTMM framework to test the extent to which subscores from the two different instruments were measuring common language achievement constructs versus aspects of student performance specific to the assessment format.

METHOD

Two instruments, the Iowa Tests of Basic Skills (ITBS) and a state developed and administered writing portfolio program (PWA), were used to explore the effects of assessment format in measuring the language abilities of elementary school students. Computerized records for the two instruments were collected and matched for all students in a large urban school district in the Southwestern United States. Students in the first sample (referred to hereafter as the 1994-95 sample) took the selected-response, ITBS Form J-Level 9 multilevel battery (Hoover, Hieronymus, Frisbie, & Dunbar, 1992) as third graders in the Spring of 1994 and a constructed-response writing portfolio assessment distributed when the students were fourth graders in the Fall of 1995. Students in the second sample (referred to hereafter as the 1996-97 sample) took the selected-response, ITBS Form J-Level 9 multilevel battery as third graders in the Spring of 1996 and began the portfolio writing assessment as fourth graders in the Fall of 1997. Students were eliminated from the sample if they had only taken one of the assessments, showed any mismatch of name or school, or were identified as special education students or students receiving modified test administration. These procedures resulted in a sample of 4,135 students for the 1994-95 sample and 4,043 students for the 1996-97 sample.

ASSESSMENT INSTRUMENTS

The ITBS and the PWA were mandated by the state at the time of administration[1]. The grade and time of administration as well as rules and procedures for test administration were dictated by state policy. There are substantial congruencies in the labels and implied constructs measured by the two assessment instruments. For example, both instruments are designed to measure language use and expression; both instruments are designed to measure language mechanics skills like spelling, capitalization, and punctuation. The method of measuring these constructs differs, however, with the writing assessment depending on the evaluation of a written, constructed-response by the student, while the ITBS is composed of selected-response items.

The Iowa Tests of Basic Skills (Hoover, et al, 1992) is a wide-range battery that measures school achievement in several content areas, including language achievement, exclusively using items with a selected-response format. The ITBS is one of the most widely used and accepted measures of student achievement (Lane, 1992; Linn, 1989). In development, construction, administration, and interpretation it represents a prototypical example of a traditional instrument for the assessment of student achievement. Six subtests of the ITBS that measure language skills were used in the present study: Vocabulary, Reading, Spelling, Capitalization, Punctuation, and Language Usage and Expression.

The New Mexico Portfolio Writing Assessment (PWA) is a state mandated assessment program first administered in the1991-92 school year and is organized and administered by the State Department of Education (NMSDE, 1997). The program is designed to provide an environment in which writing is valued and integrated into classroom activities. The state distributes three writing prompts in the fall to all teachers at participating grade levels. Students then write multiple essays drafts to all three prompts throughout the fall. Emphasis is placed on regular practice and feedback on the effective components of the writing process. Early each spring, the State Department of Education selects one of the original three prompts as the official writing prompt. Students are asked to copy their "best piece" from earlier portfolio work and this one final, essay response is collected. Operationally, only this final, capstone, constructed-response is graded by the state even though the context for the assessment is as part of a larger portfolio writing program. Therefore, for our purposes in the present study, the PWA should be viewed as a single writing assessment using constructed-response methods rather than a portfolio assessment. Rubrics are developed using benchmarks provided by a state committee which includes classroom teachers. An out-of-state contractor scores the assessment. Each response is scored on four analytic scales: Development (i.e. organization, detail, and clarity of writing); Word Usage (i.e., correct use of vocabulary and grammatical forms); Sentence Formation (i.e., correct use of sentence structure); and Language Mechanics (i.e., correct use of punctuation, capitalization, and spelling). Scores range from one to three with one representing "inadequate skill", two representing "adequate skill", and three representing "well developed skill" in the area.

CONFIRMATORY FACTOR ANALYSIS MODELS

There is some substantial discussion on the proper methods for representing and estimating MTMM models (Coenders & Saris, 2000; Dudgeon, 1994; Kenny & Kashy, 1992; Marsh, 1989; Marsh & Grayson, 1995). The models applied here are consistent with the confirmatory factor analysis (CFA) model for MTMM data (Althauser, Heberlein, & Scott, 1971; Widaman, 1985). All CFA analyses reported below employed maximum likelihood estimation methods using AMOS 4.0 (Arbuckle & Wothke, 1999). Several goodness of fit (GOF) indices were used to evaluate model fit including the chi-square test (χ^2), Adjusted Goodness of Fit Index

(AGFI), the Tucker-Lewis index (TLI), the Root Mean Square Error of Approximation (RMSEA), and the Standardized Root Mean square Residual (SRMR). As described in Hu and Bentler (1999), good model fit is indicated by AGFI and TLI values greater than .95, RMSEA values less than .06, and SRMR values less than .08.

Based on theoretical considerations and the findings of Stevens and Clauser (1995), a confirmatory factor analysis model was conceptualized based on the hypothesis that each instrument was measuring three separable, correlated language achievement constructs, but also on the hypothesis that the instruments were measuring some amount of variance that was specific to the assessment format (i.e., constructed- versus selected-response). Consistent with MTMM parlance, we often refer to the construct effects as “traits” and the assessment format effects as “methods”. This conceptualization is labeled the Correlated Traits, Correlated Methods (CTCM) model. The CTCM model represented the 10 measured variables (six ITBS and four PWA) in terms of five hypothesized latent variables: three trait factors and two method factors (see Figure 1). The three traits were Verbal Comprehension, Language Usage, and Language Mechanics. The two methods were Constructed-Response and Selected-Response. We hypothesized that the PWA Development, ITBS Reading, and ITBS Vocabulary scores would load on Verbal Comprehension. The PWA Word Usage, PWA Sentence Formation, and ITBS Use and Expression scores were loaded on the Language Usage factor. The Language Mechanics factor was related to PWA Sentence Mechanics, ITBS Spelling, ITBS Capitalization, and ITBS Punctuation. All PWA scores were loaded on the Constructed-Response method factor and all ITBS subtests were loaded on the Selected-Response method factor.

In order to evaluate competing hypotheses regarding the interrelationships among the two assessment instruments, we also tested nine additional models suggested by the taxonomy of MTMM models by Widaman (1985). The null, independence model was used as a baseline for model comparisons and was composed of 10 independent and uncorrelated factors, each representing one of the measured variables. The General Trait (GT) model tested whether a single unitary language construct fit the relationships among all measured variables. The Uncorrelated Traits (UT) model tested whether three uncorrelated language constructs fit the data and the Correlated Traits (CT) model allowed the three language constructs to be intercorrelated. The Perfectly Correlated Traits model (PCT) forced the three language constructs to be correlated at 1.0. The Uncorrelated Methods (UM) model allowed no language constructs, but specified two method of assessment factors, constructed-response and selected-response. The Correlated Methods (CM) model allowed the intercorrelation of the two methods factors.

We also tested models that combined trait and method factors. As described above, the CTCM model consisted of three correlated language constructs and two correlated method factors. The Uncorrelated Traits, Correlated Methods (UTCM) model also used three trait and two method factors, but only the method factors were allowed to correlate. The Correlated Traits, Uncorrelated Methods model (CTUM) included three correlated trait factors and added two uncorrelated method factors.

Through comparisons of particular pairs of these models, specific hypotheses about the effects of constructs and assessment format were evaluated statistically. For example, the improvement in model fit that occurs from including the language constructs in the model can be evaluated by taking the difference in chi-square for the Null model and the Correlated Traits (CT) model. Since the difference between two chi-square values for nested models is also distributed as a chi-square variable (Bollen, 1989), the significance of adding components to models can be evaluated directly through this subtraction method (see for example Byrne, 1989; Marsh, 1990; or Widaman, 1985; for further details and examples of hypothesis testing of MTMM matrices).

RESULTS

Prior to analysis, both samples were screened for outliers and missing data. No outliers were detected or excluded from the samples. Cases with missing data on any of the subscores were excluded listwise. Ninety cases with missing data were identified and excluded from the 1994-95 sample resulting in a final sample size of 4,045. For the 1996-97 sample, 110 cases with missing data were excluded resulting in an effective sample size of 3,933. Correlations, means, and standard deviations among the subscores for both assessment instruments are shown in Table 1 for the 1994-95 sample and in Table 2 for the 1996-97 sample. All correlations among the subscores in both tables were positive and significantly greater than zero ($p < .01$).

In the Campbell and Fiske terminology, there are four types of MTMM coefficients: 1) measures of the same trait by the same method (Monotrait-Monomethod Coefficients), which provide evidence of internal consistency; 2) measures of the same trait using different methods (Monotrait-Heteromethod coefficients), which provide evidence of convergent validity; and 3) measures of different traits using the same methods (Heterotrait-Monomethod coefficients) or different methods (Heterotrait-Heteromethod coefficients), which provide evidence of discriminant validity. In Tables 1 and 2, Monomethod coefficients are highlighted, Monotrait-

Heteromethod coefficients are in boldface and underlined, and the remaining coefficients are Heterotrait-Heteromethod coefficients. According to Campbell and Fiske (1959) all convergent validity coefficients (bold, underlined) should be significant and large enough to support validity. It can be seen in Tables 1 and 2 that, although the convergent validity coefficients were all positive and significantly different from zero, the magnitude of the correlations do not generally support strong convergent validity of the instruments. The mean convergent validity coefficient in Table 1 for the 1994-95 sample was .373 and in Table 2 for the 1996-97 sample was .237.

Campbell and Fiske's criteria also require that the convergent validities should be larger than correlations between measures of different traits using different methods (Heterotrait-Heteromethod), thereby establishing discriminant validity. The average Heterotrait-Heteromethod correlation for the 1994-95 sample was .368. The average Heterotrait-Heteromethod correlation for the 1996-97 sample was .226. While the convergent validity coefficients were slightly larger (.373 and .237) in magnitude than the average Heterotrait-Heteromethod correlations for both samples, the differences were not significant ($p = 0.996$).

It can also be seen that the correlations within each monomethod-heterotrait triangle (highlighted correlations) are uniformly larger than the convergent validities. The average monomethod-heterotrait correlation for the 1994-95 sample was .641 and was .565 for the 1996-97 sample. This suggests the presence of a format effect among the subscores.

In order to provide tests of additional, focused hypotheses regarding these relationships, we applied the ten CFA models described earlier. Table 3 presents the results of fitting the ten alternative MTMM CFA models. Examination of the goodness of fit indices for the ten models shows that the best fit occurred for the Correlated Traits-Correlated Methods model (CTCM), which specified three correlated trait factors and two correlated method factors. This model resulted in the largest values of the AGFI and TLI indices and the smallest values of χ^2, RMSEA, and SRMR for the 1994-95 sample. For the 1996-97 sample, the CTCM model was comparable in goodness of fit to the CTUM model and superior to all other competing models. All goodness of fit indices for the both the CTCM and CTUM model in both samples exceeded the suggestions for good model fit proposed by Hu and Bentler (1999). Comparison of the TLI indices for these two models showed a difference of less than one percent in fitted model variance. Since the CTUM model is slightly more parsimonious, it was used as the best representation of the relationships among the variables and was used for additional model comparisons and analysis. Figure 1 shows the CTUM model with standardized coefficients for the 1994-95 sample listed first and for the 1996-97 sample listed second.

Most of the remaining models displayed in Table 3 showed unacceptable model fit with AGFI and TLI values well below .90 and RMSEA and SRMR values above .10. Two exceptions were the UM and CM models, both models that fitted methods factors to the data and resulted in some degree of model fit. It should also be noted that inadmissable solutions occurred in the application of the UTCM model to both samples. Inadmissable solutions, Heywood cases and convergence problems are common difficulties in the application of CFA to MTMM matrices (Kenny & Kashy, 1992; Marsh, 1989) and usually indicate that the model being applied involves some misspecification of the relationships in the data. In the present case, results for the UTCM model can not be interpreted with any confidence.

Our primary interest in the study, however, was not in fitting a particular MTMM model, but in comparing competing MTMM models to test the significance of language constructs, assessment formats, and the relative contributions of these effects in accounting for the measured variance of the subscores. We tested eight model comparison hypotheses, two that tested the utility of assessment format factors, four that evaluated the usefulness of the language constructs, and two that examined the relative contribution of format and construct factors. Table 4 lists each hypothesis (A-H) comparing the MTMM models along with the corresponding differences in chi-square (χ^2_D), degrees of freedom (df_D), and TLI values between each pair of models tested.

As can be seen in Table 4, in comparison to the null model, a significant and large improvement in model fit occurred from inclusion of method factors that describe the different assessment formats used by the instruments (Hypothesis A). The Tucker-Lewis indices show that, for both samples, almost 90% of the subscore variance unexplained by the null model is accounted for through the inclusion of factors describing the assessment format. Although statistically significant, only modest improvements in model fit occurred by allowing the method factors to correlate with each other (Hypothesis B), suggesting that the two assessment formats were relatively independent.

The next four hypotheses tested the usefulness of construct factors in modeling the relationships among the subscores. Significant and substantial improvements in model fit occurred through the addition of trait factors (Hypothesis C) and by allowing the trait factors to correlate with each other (Hypothesis D). Three separate trait factors were superior to a single, general factor (Hypothesis E) or to perfectly correlated trait factors (Hypothesis F).

Hypothesis G compared the benefit of adding methods factors to a trait model and resulted in a significant improvement in model fit. Approximately 21% of the subscore variance in the 1994-95 sample and 17% of the subscore variance in the 1996-97 sample was fitted through the addition of the methods factors. Hypothesis H evaluated the addition of trait factors to a methods factors only model and also resulted in significant improvement in model fit. Addition of traits to the methods model resulted in a somewhat smaller improvement in fitted variance, however, approximately 11% for both samples.

Another way to evaluate the relation of instrument subscores with either format or construct effects is to compute estimates of variance components associated with the pertinent factors. Estimates of the variance associated with trait factors, method factors, and uniquenesses were obtained by taking the mean of the squared loadings for each set of factors. In these models, "unique" variance represents the reliable variance of each subscore specific to the subscore and unrelated to the rest of the CFA model as well as variance due to unreliability of measurement.

Squared loadings for the CTUM model for both samples appear in Table 5 for each subscore separately as well as for the CTUM model overall. Average squared loadings for each component were converted to percentages in the rows labeled "variance component" (see Marsh & Byrne, 1993, for further details). For the entire model, there was approximately 50% trait variance, 17% method variance, and 33% variance specific to individual subscores for the 1994-95 sample. For the 1996-97 sample, there was approximately 44% trait variance, 15% method variance, and 41% variance specific to individual subscores.

Inspection of the loadings and variance components also shows substantial differences between the two instruments. Individual subscore variances were larger for the method than the trait factors for all of the PWA subscores, but for none of the six ITBS subtests. Averaged over the two samples, the trait variance component was 14% for the PWA and 69% for the ITBS. The method variance component averaged over the two samples was 33% for the PWA and 4% for the ITBS. It is also noteworthy that the unique variance for a number of subscores was quite large in relation to the trait and method components, especially for the PWA subscores. Averaged over the two samples, the unique variance component was 52% for the PWA and 27% for the ITBS.

Discussion

The results of the present study provided evidence that, despite the common labels used to describe the measures, the two instruments assess substantially different constructs. Evaluation of convergent and discriminant validity evidence showed that relationships among measures were stronger within assessment format than within construct, across format. Using confirmatory factor analysis methods, we found that the best-fitting and most parsimonious model (CTUM) among a variety of competing models represented the data using three intercorrelated language achievement traits and two uncorrelated assessment format factors. The language achievement traits were found to be significantly different but were all highly intercorrelated.

A possible limitation in the present study was the time interval between administration of the ITBS and the PWA in each sample. The ITBS was administered in the spring and writing activity on the PWA began in the fall of the following school year. As a result, the Monotrait-Heteromethod correlations can be considered convergent validity coefficients but not concurrent validity coefficients. If there was an impact of this time interval, it would likely result in suppression of the convergent validities. To estimate the possible impact of suppression, we applied the correction for attenuation formula to the observed average convergent validity coefficients. The ITBS manual (Hoover, et al, 1992) reports test-retest reliabilities over a one-year period from grade 3 to grade 4 that average .70 over the four language subtests used in the present study. No test-retest reliability studies are available for the PWA. Assuming the same level of reliability for the PWA as the ITBS, the average convergent validity coefficients corrected for attenuation are .53 for the 1994-95 sample and .34 for the 1996-97 sample. Even if a lower test-retest reliability of .50 is assumed for the PWA, the convergent validity coefficients corrected for attenuation only rise to .63 and .40 for the two samples respectively. These estimates suggest that any confounding that may have occurred due to time of administration would not substantively affect the study conclusions. Furthermore, no effect of time interval should be expected on other comparisons in the study including the relationships among subscores within assessment method or the partitioning of variance components into common and unique variance.

Study results showed substantial differences between the two instruments in the size of different variance components in the best-fitting CFA model. For the constructed-response PWA, the majority of the variance among the subscores was associated with either the method of assessment or unique aspects of the subscores unrelated either to traits or to the method of measurement. In fact, for the PWA in both samples, the single largest variance component was unique variance. These findings reflect those of Ackerman and Smith (1988) who found unique variances greater than 60% for most variables studied. In the study by Stevens and Clauser

(1995), while there was approximately 19% unique variance, the majority of variance among subscores was method variance (55%). Results for the two samples in the present study generally show larger trait effects (47% across samples), smaller method effects (16% across samples), but larger proportions of unique variance (37% across samples).

Some aspects of the reported results also appear to reflect findings in other research that constructed-response or other alternative assessment formats may result in less stable and more task-specific measurement. Careful inspection of the results of Stevens and Clauser (1995) and the present results shows that there is greater trait and common variance associated with the selected-response subtests than with the constructed-response scores. The average Monotrait-Monomethod correlation for the PWA scores in Stevens and Clauser (1995) was .505 and in the present study was .549 for the 1994-95 sample and .381 for the 1996-97 sample. For the ITBS, the average Monotrait-Monomethod correlation in Stevens and Clauser (1995) was .613 and was .678 for the 1994-95 sample and .639 for the 1996-97 sample. In addition, while the inter-relationships among the selected-response subtests appear to be similar over the different samples, the relationships among the constructed-response scores appears to show more variability over samples. Possible influences on this may be differences in scoring or rubrics from one occasion to the next or that the prompts are not equated from year to year, difficulties of alternative assessment formats noted in other research (e.g., Camp, 1993a, 1993b; Fitzpatrick, Ercikan, Yen, & Ferrara, 1998; Green, 1995; Legg, 1998; Reckase, 1993).

Study results indicated that the two instruments are sufficiently different to warrant different score interpretations in everyday educational practice. In point of fact, a typical user of these instruments (i.e., parent, teacher, or administrator) would likely believe that the instruments can be interpreted in exactly the same way given instrument labels and descriptions (i.e., “language mechanics” from ITBS is the same as “language mechanics” from PWA). There is little evidence from our results, however, to justify the validity of trait equivalence and score interpretation as emphasized by Messick (1993; 1994). The results suggest that conclusions regarding the language achievement of these children would depend on which format was chosen for assessment. It is also difficult to speculate that the constructed-response format is measuring different but relevant language constructs when there were relatively low intercorrelations of the PWA subscores and the magnitude of unique variance was so large. In conclusion, our findings show that tests that may appear to be measuring the same constructs may in fact be measuring substantially different things when different assessment formats are used.

References

Ackerman, T. A., & Smith, P. L. (1988). A comparison of the information provided by essay, multiple-choice, and free-response writing tests. Applied Psychological Measurement, 12(2), 117-128.

Althauser, R. P., Heberlein, T. A., & Scott, R. A. (1971). A causal assessment of validity: The augmented multitrait-multimethod matrix. In H. M. Blalock (Ed.), Causal models in the social sciences (pp. 151-169). Chicago, IL: Aldine.

Arbuckle, J. L., & Wothke, W. (1999). Amos 4.0 user's guide. Chicago, IL: Smallwaters Corporation.

Baxter, G. P., Shavelson, R. J., Goldman, S. R., & Pine, J. (1992). Evaluation of procedure-based scoring for hands-on science assessment. Journal of Educational Measurement, 29, 1-17.

Bennett, R. E., Rock, D. A., Braun, H. I., Frye, D., Spohrer, J. C., & Solloway, E. (1990). The relationship of expert-system scored constrained free-response items to multiple-choice and open-ended items. Applied Psychological Measurement, 14, 151-162.

Bennett, R. E., Rock, D. A., & Wang, M. (1991). Equivalence of free-response and multiple-choice items. Journal of Educational Measurement, 28(1), 77-92.

Bollen, K. A. (1989). Structural equations with latent variables. New York: John Wiley and Sons.

Bond, L. A., Braskamp, D., van der Ploeg, A., & Roeber, E. (1996). State student assessment programs database: School year 1994-95. Oak Brook, IL: North Central Regional Educational Laboratory.

Bridgeman, B., & Rock, D. A. (1993). Relationships among multiple-choice and open-ended analytical questions. Journal of Educational Measurement, 30(4), 313-329.

Burger, S. E., & Burger, D. L. (1994). Determining the validity of performance based assessment. Educational Measurement: Issues and Practice, 13(1), 9-15.

Byrne, B. M. (1989). A Primer of LISREL. Basic Applications and Programming for Confirmatory Factor Analytic Models. New York, NY: Springer-Verlag.

Camp, R. (1993a). The place of portfolios in our changing views of writing assessment. In R. E. Bennett & W. C. Ward (Eds.), Construction versus choice in cognitive measurement (pp. 183-212). Hillsdale, NJ: Lawrence Erlbaum.

Camp, R. (1993b). Changing the model for the direct assessment of writing. In M. M. Williamson & B. A. Huot (Eds.), Validating holistic scoring for writing assessment: Theoretical and empirical foundations (pp. 45-78). Cresskill, NJ: Hampton Press.

Campbell, D. T. (1960). Recommendations for APA test standards regarding construct, trait, or discriminant validity. American Psychologist, 15, 546-553.

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

Campbell, J. R. (2000). Cognitive processes elicited by multiple-choice and constructed-response questions on an assessment of reading comprehension. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Coenders, G., & Saris, W. E. (2000). Testing additive, multiplicative, and general multitrait-multimethod matrices. Structural Equation Modeling, 7(2), 219-250.

Collins, A. (1990). Reformulating testing to measure thinking and learning. In N. Fredricksen, R. Glaser, A. Lesgold, & M. G. Shafto (Eds.), Diagnostic monitoring of skill and knowledge acquisition (pp. 75-88). Hillsdale, NJ: Erlbaum Associates.

Council of Chief State School Officers (CCSSO). (1999). Annual survey, state student assessment programs: A summary report, Fall, 1999. Washington, DC: Author.

Dudgeon, P. (1994). A reparameterization of the restricted factor analysis model for multitrait-multimethod matrices. British Journal of Mathematical and Statistical Psychology, 47, 283-308.

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. Applied Measurement in Education, 4(4), 289-303.

Fitzpatrick, A. R., Ercikan, K., Yen, W. M., & Ferrara, S. (1998). The consistency between raters scoring in different test years. Applied Measurement in Education, 11(2), 195-208.

Fredricksen, N. (1984). The real test bias. American Psychologist, 39, 193-202.

Gardner, H. (1992). Assessment in context: The alternative to standardized testing. In B. Gifford & M. C. O'Connor (Eds.), Changing assessments: Alternative views of aptitude, achievement, and instruction (pp. 77-119). Boston, MA: Kluwer Academic.

Green, B. F. (1995). Comparability of scores form performance assessments. Educational Measurement: Issues and Practice, 14(4), 13-15, 24.

Hambleton, R. K., Jaeger, R. M., Koretz, D., Linn, R.L., Millman, J., & Phillips, S. E. (1995). Review of the Measurement Quality of the Kentucky Instructional Results Information System, 1991-1994, Final Report. Frankfort, KY: Office of Educational Accountability.

Hancock, G. R. (1994). Cognitive complexity and the comparability of multiple-choice and constructed-response formats. Journal of Experimental Education, 62(2), 143-157.

Hoover, H. D., Hieronymus, A. N., Frisbie, D. A., & Dunbar, S. B. (1992). Iowa tests of basic skills. Chicago, IL: The Riverside Publishing Company.

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

Katz, I. R., Bennett, R. E., & Berger, A. E. (2000). Effects of response format on difficulty of SAT-Mathematics items: It's not the strategy. Journal of Educational Measurement, 37(1), 39-57.

Kenny, D. A., & Kashy, D. A. (1992). Analysis of the multitrait-multimethod matrix by confirmatory factor analysis. Psychological Bulletin, 112(1), 165-172.

Koretz, D., Stecher, B. M., Klein, S., & McCaffrey, D. (1994). The Vermont portfolio assessment program: Findings and implications. Educational Measurement: Issues and Practice, 13(3), 5-16.

Lane, S. (1992). Review of the Iowa Tests of Basic Skills, Form J. In J. J. Kramer & J. C. Conoley (Eds.), The Eleventh Mental Measurements Yearbook (pp. 421-423). Lincoln, NE: The University of Nebraska Press.

Legg, S. M. (1998). Reliability and validity. In W. Wolcott & Sue M. Legg (Eds.), An overview of writing assessment: Theory, research, and practice (pp. 124-142). Urbana, IL: National Council of Teachers of English.

Linn, R. L. (1989). Review of the Iowa Tests of Basic Skills, Forms G and H. In J. C. Conoley & J. J. Kramer (Eds.), The Tenth Mental Measurements Yearbook (pp. 393-395). Lincoln, NE: The University of Nebraska Press.

Lu, C., & Suen, H. K. (1995). Assessment approaches and cognitive style. Journal of Educational Measurement, 32(1), 1-17.

Lukhele, R., Thissen, D., & Wainer, H. (1994). On the relative value of multiple-choice, constructed-response, and examinee-selected items on two achievement tests. Journal of Educational Measurement, 31(3), 234-250.

Marsh, H. W. (1989). Confirmatory factor analysis of multitrait-multimethod data: Many problems and few solutions. Applied Psychological Measurement, 13, 335-361.

Marsh, H. W. (1990). Confirmatory factor analysis of multitrait-multimethod data: The Construct validation of multidimensional self-concept responses. Journal of Personality, 58, 661-692.

Marsh, H. W., Byrne, B. M. (1993). Confirmatory factor analysis of multitrait-multimethod self-concept data: Between-group and within-group invariance constraints. Multivariate Behavioral Research, 28, 313-349.

Marsh, H. W., & Grayson, D. (1995). Latent variable models of multitrait-multimethod data. In R. H. Hoyle (Ed.), Structural equation modeling: Concepts, issues, and applications. Thousand Oaks, CA: Sage.

Martinez, M. E., Katz, I. R. (1996). Cognitive processing requirements of constructed figural response and multiple-choice in architecture assessment. Educational Assessment, 3(1), 83-98.

Messick, S. (1993). Trait equivalence as construct validity of score interpretation across multiple methods of measurement. In R. E. Bennett & W. C. Ward (Eds.), Construction versus choice in cognitive measurement (pp. 61-74). Hillsdale, NJ: Lawrence Erlbaum.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. Educational Researcher, 23(2), 13-23.

Moss, P. (1994). Can there be validity without reliability? Educational Researcher, 23(2), 5-12.

New Mexico State Department of Education (NMSDE). (1997). Interpretive guide for the New Mexico Portfolio Writing Assessment program. Santa Fe, NM: Author.

Nickerson, R. S. (1989). New directions in educational assessment. Educational Researcher, 18 (9), 3-7.

Reckase, M. D. (1993). Portfolio assessment: A theoretical prediction of measurement properties. Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford and M. C. O'Connor (Eds.), Changing assessments: Alternative views of aptitude, achievement, and instruction (pp. 38-75). Boston, MA: Kluwer Academic Publishers.

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. Journal of Educational Measurement, 30, 215-232.

Stecher, B. M., Klein, S. P., Solano-Flores, G., McCaffrey, D., Robyn, A., Shavelson, R. J., & Haertel, E. (2000). The effects of content, format, and inquiry level on science performance assessment scores. Applied Measurement in Education, 13(2), 139-160.

Stevens, J. J., & Clauser, P. (1995). Multitrait-multimethod comparisons of a writing portfolio and the ITBS. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Thissen, D., Wainer, H., & Wang, X-B. (1994). Are tests comprising both multiple-choice and free-response items necessarily less unidimensional than multiple-choice tests? An analysis of two tests. Journal of Educational Measurement, 31(2), 113-123.

Traub, R. (1993). On the equivalence of the traits assessed by multiple-choice and constructed-response tests. In R. E. Bennett & W. C. Ward (Eds.), Construction versus choice in cognitive measurement (pp. 29-44). Hillsdale, NJ: Lawrence Erlbaum.

Ward, W. C. (1987). A comparison of free-response and multiple-choice forms of verbal aptitude tests. Applied Psychological Measurement, 6(1), 1-11.

Ward, W. C., Dupree, D., & Carlson, S. B. (1987). A comparison of free-response and multiple-choice questions in the assessment of reading comprehension (RR 87-20). Princeton, NJ: Educational Testing Service.

Werts, C. E., Breland, H. M., Grandy, J., & Rock, D. A. (1980). Using longitudinal data to estimate reliability in the presence of correlated errors of measurement. Educational and Psychological Measurement, 40(1), 19-29.

Widaman, K. F. (1985). Hierarchically nested covariance structure models for multitrait-multimethod data. Applied Psychological Measurement, 9, 1-26.

Wolf, D., Bixby, J., Glenn, J., & Gardner, H. (1991). To use their minds well: Investigating new forms of student assessment. Review of Research in Education, 17, 31-73.

Footnotes

[1] Since that time, the state has adopted a different achievement test and has modified the writing assessment from a portfolio format to a direct writing assessment.

Table 1

Correlations, Means (M), and Standard Deviations (SD) Among the Writing and ITBS Subscores For the 1994-95 Sample (N = 4045)

	Writing				ITBS							
Subscore	1	2	3	4	5	6	7	8	9	10	M	SD
Writing:												
1. Development	1.000										2.15	.57
2. Mechanics	.474	1.000									2.39	.63
3. Word Usage	.600	.482	1.000								2.22	.48
4. Sentence Formation	.577	.584	.579	1.000							2.20	.65
ITBS:												
5. Vocabulary	**.364**	.367	.396	.402	1.000						106.02	13.89
6. Reading	**.355**	.370	.390	.399	.800	1.000					106.10	17.60
7. Spelling	.350	**.378**	.362	.374	.663	.669	1.000				104.17	18.35
8. Capitalization	.340	**.359**	.349	.385	.619	.655	.648	1.000			106.99	17.70
9. Punctuation	.325	**.355**	.334	.384	.594	.643	.633	.740	1.000		106.05	17.64
10. Use and Expression	.362	.373	**.389**	**.414**	.730	.756	.670	.671	.684	1.000	106.26	18.71

Note. Convergent validity coefficients (Monotrait-Heteromethod) are bold, underlined; Monomethod coefficients are highlighted; the remaining correlations are Heterotrait-Heteromethod coefficients.

Table 2

Correlations, Means (M), and Standard Deviations (SD) Among the Writing and ITBS Subscores For the 1996-97 Sample (N = 3933)

	Writing				ITBS							
Subscore	1	2	3	4	5	6	7	8	9	10	M	SD
Writing:												
1. Development	1.000										2.05	.46
2. Mechanics	.387	1.000									2.18	.42
3. Word Usage	.323	.322	1.000								2.05	.24
4. Sentence Formation	.407	.517	.329	1.000							2.20	.48
ITBS:												
5. Vocabulary	**.213**	.237	.176	.286	1.000						106.71	13.73
6. Reading	**.222**	.240	.164	.282	.749	1.000					106.20	16.64
7. Spelling	.220	**.262**	.179	.283	.607	.584	1.000				104.49	17.61
8. Capitalization	.222	**.239**	.166	.278	.568	.597	.610	1.000			106.63	16.81
9. Punctuation	.206	**.229**	.166	.255	.568	.586	.619	.730	1.000		105.74	16.88
10. Use and Expression	.230	.251	**.187**	**.309**	.737	.728	.627	.636	.642	1.000	106.31	17.98

Note. Convergent validity coefficients (Monotrait-Heteromethod) are bold, underlined; Monomethod coefficients are highlighted; the remaining correlations are Heterotrait-Heteromethod coefficients.

Table 3

Model Results For The 1994-95 and the 1996-97 Samples

Model and Sample	χ^2	df	AGFI	TLI	RMSEA	SRMR
Null, Independence Model						
1994-95	25536	45	.136	--	.374	.468
1996-97	19231	45	.234	--	.329	.391
General Trait Model (GT)						
1994-95	4541	35	.653	.773	.178	.107
1996-97	3180	35	.744	.789	.151	.101
Uncorrelated Traits (UT)						
1994-95	11699	35	.472	.412	.287	.399
1996-97	9348	35	.520	.376	.260	.331
Correlated Traits (CT)						
1994-95	4000	32	.655	.781	.175	.111
1996-97	2407	32	.778	.826	.137	.102
Perfectly Correlated Traits (PCT)						
1994-95	5628	35	.588	.718	.199	.417
1996-97	5165	35	.660	.656	.193	.474
Uncorrelated Methods (UM)						
1994-95	2474	35	.834	.877	.131	.246
1996-97	1728	35	.867	.887	.111	.157
Correlated Methods (CM)						
1994-95	1142	34	.905	.942	.090	.029
1996-97	1111	34	.903	.926	.090	.031
Uncorrelated Traits-Correlated Methods (UTCM)						
1994-95	225	24	.975	.985	.046	.018*
1996-97	108	24	.988	.992	.030	.015*
Correlated Traits-Uncorrelated Methods (CTUM)						
1994-95	169	22	.979	.988	.041	.013
1996-97	72	22	.991	.995	.024	.012

Table 3 Continued

Model and Sample	χ^2	df	AGFI	TLI	RMSEA	SRMR
Correlated Traits-Correlated Methods (CTCM)						
1994-95	149	21	.981	.989	.039	.017
1996-97	72	21	.990	.994	.025	.013

Note. AGFI = Adjusted Goodness of Fit Index, TLI = Tucker-Lewis Index, RMSEA = Root Mean Square Error of Approximation, SRMR = Standardized Root Mean Square Residual.

* Application of the Uncorrelated Traits-Correlated Methods (UTCM) model to both samples resulted in inadmissable solutions.

Table 4

Tests of Hypotheses Through Comparisons of the MTMM Models

Hypothesis	χ^2_D	df_D	TLI_D
Format Effects			
A. Methods (Null - UM)			
1994-95	23,076	10	.877
1996-97	17,503	10	.887
B. Correlated Method Factors (UM - CM)			
1994-95	1,332	1	.065
1996-97	617	1	.039
Construct Effects			
C. Traits (Null - UT)			
1994-95	13,837	10	.412
1996-97	9,883	10	.376
D. Correlated Trait Factors (UT - CT)			
1994-95	7,699	3	.369
1996-97	6,941	3	.450
E. One Factor versus Three Traits (GT - CT)			
1994-95	541	3	.008
1996-97	773	3	.037
F. Correlated vs. Perfectly Correlated Traits (PCT - CT)			
1994-95	1628	3	.063
1996-97	2758	3	.170
Relative Importance of Format and Construct Effects			
G. Traits only vs. Traits + Methods (CT - CTUM)			
1994-95	3,831	10	.207
1996-97	2,335	10	.169
H. Methods only vs. Traits + Methods (UM - CTUM)			
1994-95	2,305	13	.111
1996-97	1,039	13	.108

Note. All listed chi-square values are significant, $p < .001$.

Table 5

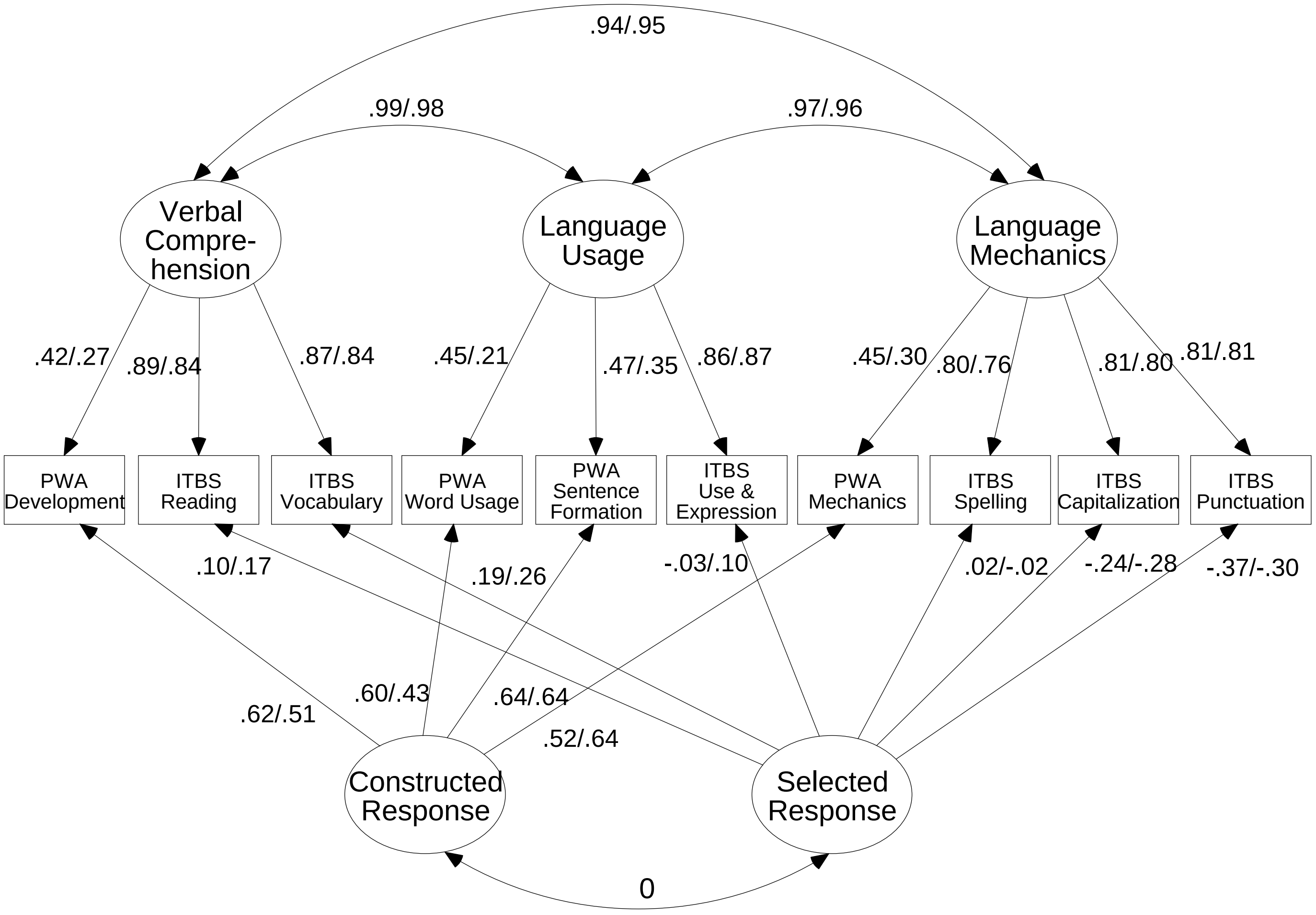
Trait, Method, and Unique Variance by Subscore for the Two Samples

	1994-95 Sample			1996-97 Sample		
Subscore	Trait	Method	Unique	Trait	Method	Unique
Writing Assessment:						
1. Development	.176	.384	.439	.073	.257	.670
2. Mechanics	.199	.272	.528	.090	.404	.505
3. Word Usage	.200	.362	.437	.046	.182	.772
4. Sentence Formation	.222	.407	.371	.124	.403	.473
PWA Variance Component	19.9	35.6	44.4	8.3	31.2	60.5
ITBS:						
5. Vocabulary	.762	.037	.201	.709	.070	.222
6. Reading	.796	.010	.195	.697	.029	.274
7. Spelling	.642	.001	.357	.573	.001	.426
8. Capitalization	.651	.056	.292	.642	.077	.281
9. Punctuation	.650	.140	.210	.650	.092	.259
10. Usage and Expression	.733	.001	.266	.757	.009	.233
ITBS Variance Component	70.6	4.1	25.4	67.1	4.6	28.3
Total Variance Component	50.3	16.7	33.0	43.6	15.2	41.2

Note. The reported variance component is the mean of the squared factor loadings in the column expressed as a percentage.

Figure Caption

Figure 1. The correlated traits-uncorrelated methods (CTUM) model for the 1994-95 and 1996-97 samples.


.94/.95
.99/.98
.97/.96
Verbal Compre-hension
Language Usage
Language Mechanics
.42/.27
.89/.84
.87/.84
.45/.21
.47/.35
.86/.87
.45/.30
.80/.76
.81/.80
.81/.81
PWA Development
ITBS Reading
ITBS Vocabulary
PWA Word Usage
PWA Sentence Formation
ITBS Use & Expression
PWA Mechanics
ITBS Spelling
ITBS Capitalization
ITBS Punctuation
.10/.17
.19/.26
-.03/.10
.02/-.02
-.24/-.28
-.37/-.30
.62/.51
.60/.43
.64/.64
.52/.64
Constructed Response
Selected Response
0