Application of Confirmatory Factor Analysis to the Validity Study of a Performance Assessment:
A Multitrait-multimethod Structure and its Invariance across Gender and Grade

Dong-il Kim*

Seoul National University

Abstract

This study investigated construct validity and factorial invariance of a performance assessment of reading comprehension and writing proficiency, through a multitrait-multimethod structure (MTMM), using the confirmatory analysis technique. First, interrater reliability was examined for each measured variable using three different generalizability coefficients. Although all of the measures were found to be highly reliable, exploratory factor analysis indicated that trait and method effects were confounded in the measured variables. Consequently, confirmatory factor analysis was used to disentangle multidimensionality and examine the convergent and discriminant validity of the latent variables according to the Campbell-Fiske criteria. These analyses indicated that a model with three correlated trait factors and three correlated method factors provided the best fit to the data. Finally, a factorial invariance across gender and grade was examined. While this MTMM factorstructure was fitted to the data in each subgroup (fifth grade boys. fifth grade girls, sixth grade girls), the factorial invariance across gender and grade was supported only in a particular set of parameters. Methodological and practical implications of the use of confirmatory factor analysis in multitrait-multimethod analyses are also discussed for construct validation in performance assessment across different groups.

Key words: performance assessment, multitrait-multimethod structure, confirmatory factor analysis, factorial invariance, construct validation

Performance assessment generally refers to a task (problem) that requires an individual to actively construct a response (solution), as opposed to simply recalling memorized knowledge (Baron, 1991). Although performance assessment has been quite popular in such areas as administration and management (Berk, 1986;Priestley, 1982), mechanical job performance appraisal (Priestley, 1982) and teacher evaluation (Stiggins & Bridgeford, 1984), it is only recently that performance assessment has been considered a

* Corresponding author Tel: +82-2-880-7636
E-mail address: dikimedu@snu.ac.kr

viable approach to large scale testing of students academic achievement (Kim, 1992).

If performance assessment is to be an acceptable alternative to traditional multiple-choice tests, it must be publicly accountable and professionally credible; that is, it must show sound technical adequacy with respect to reliability, validity, and scoring procedures (American Educational Research Association, American Psychological Association, and the National Councilon Measurement in Education, 1985). Sometimes, however, these psychometric properties seem to be difficult to achieve with performance measures (Mehrens, 1992). An objective and reliable scoring of performance assessments requires careful and systematic training for examiners, which can be both time-consuming and expensive. Furthermore, performance assessments often have no evidence of validity other than face validity. Some degree of face validity may be essential for public acceptance, but this is not sufficient as the sole indicator of validity, particularly when the assessments are used in "high stakes" testing programs.

Questions concerning whether a test measures what it is intended to measure are answered through assessment of construct validity. Construct validity integrates a theoretical rationale with empirical evidence that bears on the interpretation of meaning of a measure (Messick, 1989). A construct can be defined as a product of informed scientific imagination an idea developed to permit categorization and description of some directly observable behavior (Crocker & Algina, 1986). Traditionally, construct validation evidence is assembled through a series of studies including experimental, correlational, and discriminant approaches. When the adequacy of the test as an indicator of a construct is of primary concern, exploratory factor analysis and internal consistency assessment are typically conducted.

Compared to multiple-choice tests, the construct validation of performance assessments using constructed-response poses some additional problems. Regardless of the domain of assessment, language abilities, in particular, are likely to significantly influence scores [2], because most performance assessment requires students to demonstrate knowledge by actively constructing written response, subject matter scores will be confounded with language skills. For example, students' written responses to open-ended mathematical problems will be influenced not only by their understanding of mathematics, but by their language fluency and writing abilities as well. More generally, "constructs" and "items"(questions) are likely to be confounded in performance assessment because multiple constructs are likely to be embedded in each item. Consequently, the relevant construct and the irrelevant method effects are entangled, and a unidimensional approach such as exploratory factor analysis fails to provide an adequate examination of construct validity; instead, a multidimensional analysis is required.

This thread to the validity of performance assessment implies a potentially adverse impact on certain population subgroups. Because the response requires multiple traits,

2) Of course, reading ability influence scores on multiple-choice tests as well, but scores from performance assessments are influenced by expressive language abilities in addition to reading abilities.

it is not easy to just measure the target component. One of the well-documented areas is a gender difference in performance assessment (Bennett, 1993). Several studies have found that relative to boys, girls perform better on constructed-response that on multiple-choice items. These gender-related format differences can be hypothetically explained thus: girls may perform better because the constructed-response requires some construct-irrelevant attributes in which girls are strong (i.e., writing proficiency and verbal ability).

Research on gender difference in intellectual abilities has long been of interest to educators, which has found that girls tend to score higher than boys on tests of language usage (spelling, grammar) and perceptual speed (Feingold, 1992). Contemporary investigations have focused on two aspects: (a) difference in average performance through the meta-analytic review (Born, Bleichrodt, & Van Der Flier, 1987; Hyde & Linn, 1988)ortheanalysis of norms from standardized tests (Martin & Hoover, 1987) and (b) difference in variability in intellectual abilities (Feingold, 1992). In terms of psychometric studies on performance assessment, gender difference in mean levels of test scores is not necessarily a test bias. This difference may accurately represent essential distinctions in group performance. Additionally, trend analyses have revealed that gender differences in intellectual abilities among adolescents have decreased markedly over the past generation (Feingold, 1988; Jacklin, 1989). As for performance assessment, the results of a recent state-wide alternative assessment system in Minnesota, using constructed-response, showed that boys seemed catch up with girls in junior high school level and score even better at the high school level,eventhough girls did better in elementary level, in general (M. Davison, personal communication, April, 1994).

Therefore, a more fundamental issue about construct validity is whether responses to the same test have the same meaning for boys and girls.

One classical approach to multidimensional analysis on construct validity is the multitrait-mulitimethod (MTMM) matrix developed by Campbell and Fiske (1959). With this technique, not only the constructs of interest but other dimensions of measurement (method effects) are explicitly considered. An MTMM matrix is a matrix of correlations among measures of multiple traits, each of which is assessed by multiple methods. Although the MTMM matrix is the most widely used approach to evaluating multitrait-multimethoddata, this approach has been criticized because it is based on the observed correlations between measured variables. A more advanced technique is the use of confirmatory factor analysis (CFA), inferring trait and method effects on the basis of latent variables (Marsh, 1993; Marsh & Richard, 1985; Widaman, 1985; Wothke & Browne, 1990). The logic and heuristic value of the Campbell-Fiske criteria are still applicable; the difference is that they are applied to relationships among latent constructs, rather than measured variables (Marsh, 1989). Furthermore, by fixing or constraining various parameters, CFA can be use to test a variety of assumptions about the data (e.g., number of traits represented, whether traits are correlated) by specifying

different models and empirically comparing how well these alternative models fit the data. This analytic approach thus provides a much stronger basis for analyzing multitrait-multimethod data.

The purpose of this study was to explore the utility of MTMM approaches to the investigation of the construct validity of performance assessments, using the particular example of an assessment of reading comprehension and writing ability. Assessment of these abilities using constructed response measures seemed particularly challenging. Conceptually, although both reading and writing are linguistic abilities, comprehension of a passage of text is somewhat distinct from the ability to communicate this understanding to others. In practice, however, scores for comprehension and writing ability based on the same sample of writing are almost certain to be confounded to some degree. Also, because scores from performance assessments of writing ability have been found to vary greatly as a function of topic (e.g., Breland, Camp, Jones, Morris & Rock, 1987), method (question) effects are likely to be present in the data as well (i.e., scores for different traits assessed from responses to the same question may be correlated as highly as scores for the same trait assessed from the responses to different questions). Both of these factors should make it difficult to assess convergent and discriminant validity from correlations based on the measured variables. Once the MTMM structure was identified, testing for factorial invariance over different subpopulations was implemented. More specifically, we investigated whether this particular test has the same meaning for boys and girls of different grade levels.

Ⅰ. Method

A. Subjects

Students participating in this research were part of a larger, longitudinal study of children's social, ethical, and intellectual development being conducted in six school districts-three in large cites, one in a small city and two in suburban communities in the United States. The districts are geographically diverse: three on the West Coast, one in the South, one in the Southeast, and one in the Northeast. Students from four elementary schools in each of the six districts took part in the study. The performance assessment was administered to 1,023 students (46% male, 54% female) in 5th or 6th grades (Grade 5 = 57%, Grade 6 = 43%) near the end of the school year (May).

B. Assessment Instrument and Procedures

The reading comprehension assessment used a 375-word passage from "The Little Prince" (de Saint-Exupery, 1943), with a Flesch grade level of 5.3. The passage describes the prince's encounter with a fox, during which the fox expresses the view that humans are only interested in hunting and raising chickens, and defines "tameness" as a unique

bond between himself and a human being.

Student read the passage and then responded in writing to the following three questions about its meaning, under untimed conditions: (a) What did the fox mean about being tame? (b) Why does the fox want to be tame? (c) Why does the fox think men are only interested in hunting and raising chickens?

The scoring procedures were adapted from those used in the National Assessment of Educational Progress of reading and literature (National Assessment of Educational Progress, 1984), developed by the Educational Testing Service. Two trained raters scored students' written responses to the questions for Understanding (6points), Complexity of Writing (5 points), Clarity of Thought (4 points), and Grammatical Usage and Spelling (4 points). The scorers also counted the Number of Words written in response to each question. The final scales were created by averaging the two raters' scores. Because the first two questions both concerned students' understanding of the meaning of "tameness" in the passage, the first Adequacy of Understanding score was based on the written answers to both question and 2. All other measures were scored from the response to each of the three questions. Thus, there were a total of 14 scores derived from each student's responses to the three questions.Thedetailed scoring guidelines are provided elsewhere (Developmental Studies Center, 1993).

C. Analysis

Interrater reliability was investigated through generalizability theory (Shavelson & Webb, 1991). To examine construct validity, an exploratory factor analysis using oblique rotation was first performed to examine preliminary factor structure. We then conducted confirmatory factor analysis of the latent constructs using EQS (Bentler, 1989): Finally, we examined factorial invariance across gender and grade through subsequent hierarchical nested models with various constraints.

Ⅱ. Results

A. Preliminary analyses

Demonstrating that the measured variables are reliable is necessary before assessing construct validity. Because each variable was rated by two raters, of critical importance was the extent to which the scores of the two raters agreed (i.e., interrater agreement). Three generalizability (G) coefficients are reported in Table 1. The first G coefficient represents the extent to which raters rank ordered students in the same way (relative agreement). This is equivalent to the intraclass correlation coefficient. The second G coefficient, on the other hand represents the extent to which students received identical scores from the two raters (absolute agreement). In terms of technical adequacy, absolute interrater agreement coefficients of .60 and higher are considered acceptable

(Davison, 1989). Using this criterion, the level of absolute interrater agreement on every measured variable was good to excellent (.70-.99). This finding confirms that a performance assessment can be reliable with careful rater-training and appropriate scoring criteria. Finally, the third G coefficient is the reliability when both raters' scores are combined (Coefficient Alpha), which is relevant is this investigation because we crated the scale score by averaging two raters' scores. All of the measured variables used in the analyses seemed to be very reliable (.83-.99)

Conceptually, the data should represent three traits: Reading Comprehension, Writing Quality, and Writing Fluency. An exploratory factor analysis of the 14 measured variables identified three factors, as shown in Table 2. However, the factor structure did not clearly reveal the expected three traits. Factor II does appear to represent Writing Fluency, with all six of the scores for Number of Words and Complexity of Writing having their highest loadings on this factor. In Factors I and III, however, method and trait effects appear confounded. The scores for Clarity of Thought, Grammar (Grammatical usage and Spelling) and Understanding were clustered within different methods (questions)onthree factors, with scores for questions 1 and 2 having their highest loadings on the first factor, and scores for question 3 having their highest loadings on the third.

Table 1.

Interraterreliability:GeneralizabilityCoefficients

Measured Variables	Variance Component (Student)	Variance Component (Rater)	Variance Component (Interaction)	Relative Error Variance	Absolute Error Variance	G1(ICC): Relative Agreement	G2 Absolute Agreement	G3: (Alpha Coeff.)
Understanding(Q1&Q2)	0.319	0.000	0.131	0.131	0.131	0.709	0.709	0.830
Understanding(Q3)	1.216	0.000	0.209	0.209	0.209	0.854	0.854	0.921
Q1ComplexityofWriting	1.018	0.000	0.197	0.197	0.197	0.838	0.838	0.912
Q1Clarity of Thought	0.691	0.000	0.293	0.293	0.293	0.702	0.702	0.825
Q1Grammar	0.539	0.000	0.208	0.208	0.208	0.722	0.721	0.838
Q2ComplexityofWriting	0.696	0.000	0.238	0.238	0.238	0.745	0.745	0.854
Q2Clarity of Thought	0.531	0.002	0.216	0.216	0.218	0.711	0.709	0.831
Q2Grammar	0.446	0.001	0.101	0.101	0.102	0.816	0.814	0.898
Q3ComplexityofWriting	0.746	0.006	0.248	0.248	0.254	0.750	0.746	0.857
Q3Clarity of Thought	0.817	0.000	0.146	0.254	0.750	0.848	0.848	0.918
Q3Grammar	0.575	0.001	0.110	0.110	0.111	0.840	0.838	0.918
Q1No.ofWords	119.022	0.000	0.029	0.029	0.029	0.999	0.999	0.999
Q2No.ofWords	69.608	0.000	0.947	0.947	0.947	0.987	0.987	0.993
Q3No.ofWords	53.349	0.001	0.132	0.132	0.132	0.998	0.998	0.999

Notes.G1andG2arereliabilityestimates ofarandomlyselectedrating.G3 isareliabilityestimateoftworatingscombined.

Table 2. Exploratory Factor Analysis: Oblique Factor Model

Measured Variables	Factor I	Factor II	Factor III
Q1ClarityofThought	0.802	-0.085	0.076
Q2ClarityofThought	0.725	-0.054	0.116
Q1Grammar	0.558	0.054	0.141
Q2Grammar	0.69	0.082	0.285
Understanding(Q1 andQ2)	0.69	0.082	0.285
Q2No.ofWords	0.07	0.807	-0.019
Q2ComplexityofWriting	0.089	0.791	-0.072
Q3No.ofWords	-0.179	0.771	0.309
Q3ComplexityofWriting	-0.219	0.723	0.314
Q1No.ofWords	0.404	0.642	-0.197
Q1ComplexityofWriting	0.491	0.566	-0.244
Q3ClarityofThought	0.211	-0.081	0.802
Understanding(Q3)	0.161	0.028	0.779
Q3Grammar	0.074	0.219	0.55
Factorpatterncorrelations			
Factor I	1.000		
Factor II	0.358	1.000	
Factor III	0.252	0.284	1.000
Eigen Values	5.429	1.632	1.400

B. Establishing an MTMM structure using Confirmatory Factor Analysis (CFA)

MTMM analysis produces factors corresponding to the traits and methods (questions). That is, factors defined by multiple indicators derived from the same method represent method effects. MTMM analysis can be viewed as an application of confirmatory factor analysis with pre-determined factors assigned to traits and methods. An "anchor model"representing three (correlated) traits and three (correlated) method factors (corresponding to the three questions), shown in Figure 1, was fitted to the data.

An advantage of MTMM studies using confirmatory factor analysis is that a series of alternative models can be tested against the anchor model. When the identified model is able to fit the data, various parameters in the model can be constrained to generate nested models, and there alternative models can be examined for theirrelative ability to fit the data. Several criteria were used to evaluate the adequacy of the anchor model, and various alternative models, as shown in Table 3.

First, overall chi-square tests of goodness of fit, based on differences between the original and reproduced covariance matrices, are shown. This goodness of fittest, however, is dependent on sample size. Even a model which fits the data very well may produce a statistically significant chi-square for large sample sizes (Bollen & Long, 1993), as in the present case. To overcome this shortcoming, two alternative indices were considered.

Bentler and Bonett (1980) suggest that the goodness of fit of a particular model may be usefully assessed using the Comparative Fit index which has the advantage of reflecting fit relatively well at all sample sizes. The second fit criterion has been derived on the basis of information theory considerations by Akaike (1989). In the spirit of parsimony, Akaike argued that when selecting a model from a large number of models, one should take into account both statistical goodness of fit and the number of parameters that have to be estimated to achieve that degree of fit. The Akaike Information Criterion (AIC) is designed to balance these two aspects of model fit. In general, small AICs result from models with few estimated parameters and a good fit to the data, whereas models with many parameters to be estimated yield large AICs.

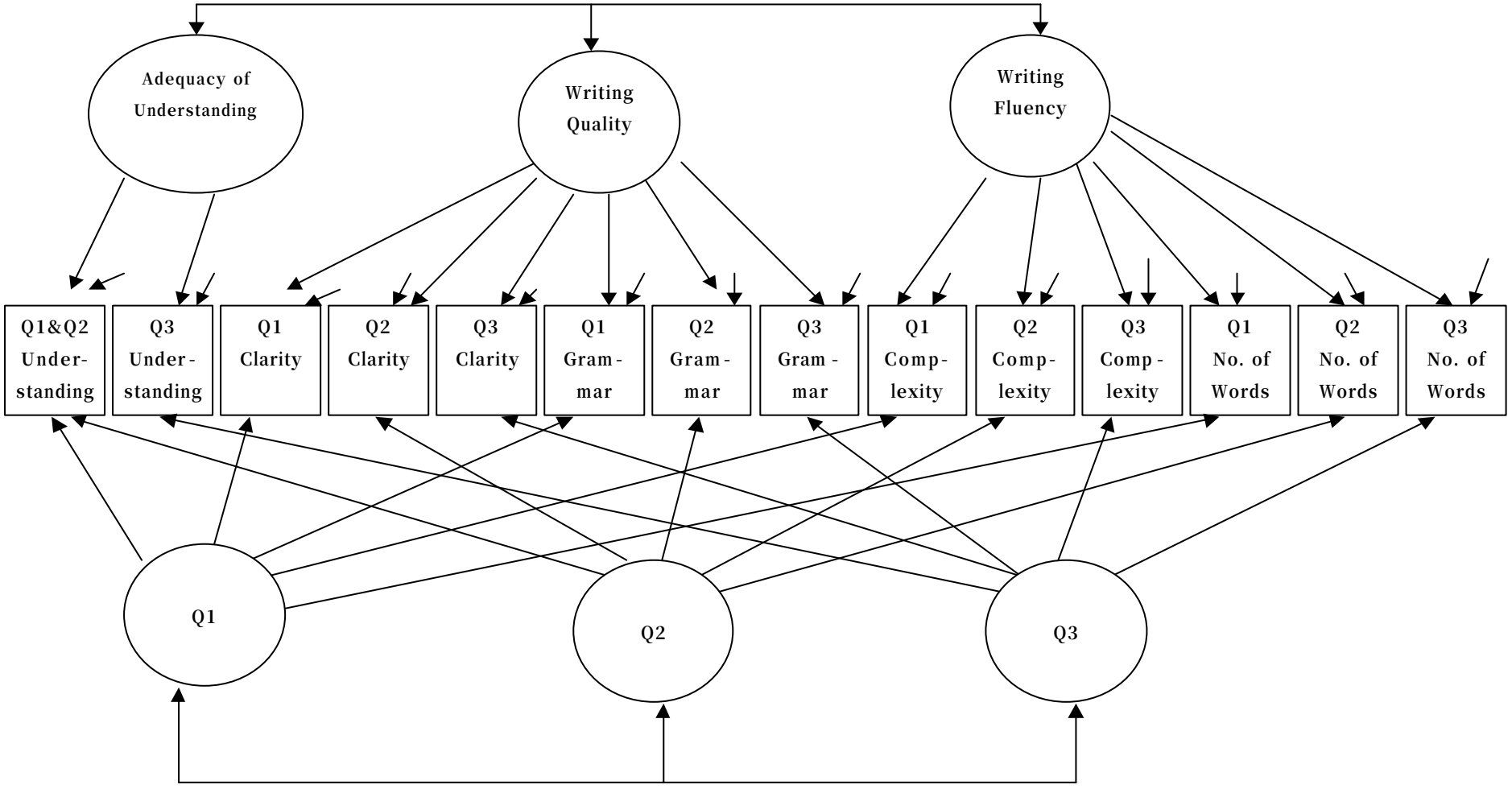

Figure 1. An anchor model (three correlated traits and three correlated methods) of MTMM structure using confirmatory factor analysis.

Table 3
Summary ofTestedModels

Modeldescriptions	2(df)	CFI	AIC	2 diff(df)
1.Anchor Model-3correlatedtrait-factors,3 correlated method-factors	760.69(56)**	.901	648.69	
2.Threecorrelatedmethod-factorswithout trait-factors	2090.76(73)**	.714	1944.76	1330.07(17)**
3.Threecorrelatedtrait-factors withoutmethod-factors	2523.28(74)**	.653	2375.28	1762.59(18)**
4.ThreeUncorrelatedtrait-factors and three correlated method-factors	958.23(59)**	.873	840.23	197.54 (3)**
5.Threecorrelatedtrait-factors and three Uncorrelated method-factors	935.16(59)**	.876	817.16	174.47 (3)**
6.Anchor Modelwith equal correlations among method-factors	780.20(58)**	.898	664.19	19.51 (2)**
7.Anchor Modelwith equal correlations among trait-factors	958.23(59)**	.873	840.23	197.54 (3)**
8.Twocorrelatedtrait-factors(UnderstandingandWritingQuality combined) and three correlated method-factors	772.04(58)**	.899	656.04	11.35 (2)*

Notes. CFI (Comparative Fit Index) is based on a Null Model with 2(df)=7149.95 (91). The 2 difference is based on the difference between the Anchor Model (Model1)andthemodelbeingtested.

* p < 0.01
**p<0.001

Although the chi-square for the three trait, three method anchor model was statistically significant due to the large sample size, CFI indicated a good fit to the data, reaching .90 or higher (Bentler, 1989). Once this anchor model is established, alternative models can be fit to the data to test various hypotheses related to the Campbell-Fiske criteria (Campbell & Fiske, 1959). These alternative models can be compared for goodness of fit by taking the differences in their chi-square values and testing against the difference in the degrees in the degrees of freedom (Bentler & Bonett, 1980). Various alternative models were assessed in the present study, and their fir indices are also summarized in Table 3.

Models 2 and 3 investigated the relative importance of method and trait factors. Model 2, including three method factors without traits, provided a poor fit to the data (CFI=.714). Model 3, containing three correlated trait factors without method factors, also showed a poor fit to the data (CFI=.653). These results indicate that both trait and method effects were necessary to adequately represent the data. The next two models therefore included both trait and method factors, but tested assumptions about the relationships traits and methods. Both Model 4, in which the traits were assumed to be uncorrelated, and Model 5, in which the method factors were assumed to be uncorrelated, provided poor fits to the data (Model 4: CFI=.873; Model 5: CFI=.876). Thus, both correlated trait factors and correlated method factors were necessary assumptions.

We next examined the question of whether the correlations among the trait and method factors could be assumed to be equal. Model6, with equal correlation of the method factors, seemed to fit the data almost as well as the anchor model (CFI=.898). However, the difference in chi-squares between the anchor model and Model 6 was highly significant. Model 7, representing equal correlation of the trait factors, provided a poor fit to the data (CFI=.873).

Finally, we examined whether a model with only two, rather than three traits, would adequately fit the data. Specifically, since the latent traits Adequacy of Understanding and Writing Quality seemed to be close each other in the exploratory factor analysis (see Table 2), the consequence of combining these two traits was examined. Although this two-trait, three methods factor model does not have a good conceptual justification, this modelprovides a test of the discriminant validity of the three trait factors. Model 8 had an acceptable fit to the data (CFI=.899), but, again, the difference in chi-squares between it and the anchor model was highly significant. In addition to the subsequent significant chi-square difference, the anchor model also had the smallest AIC value among the tested models, indicating that it was the most parsimonious model.

To summarize, the findings indicated:

1. The three trait factors were very important, showing good convergent validity, but a substantial portion of variance also depended on the method factors.
2. The three traits were significantly intercorrelated.

3. Elimination of any trait factors resulted in a significantly poorer fit. That is, discriminant validity was demonstrated in the analyses.

C. Invariance Constraints Across All Groups

The factor structure identified so far was based on data from the total sample of students. To examine the question of whether this structure would hold across four subgroups, the three-trait, three-method model was fit separately to data from boys and girls in grade 5 and 6. All four models showed an acceptable fit to the data. These results provide a support for the anchor model but do not explain the invariance of the parameter estimates across gender and grade. In order to test the appropriateness of the invariance, the hierarchical models for all four groups were also provided. The first model is the model in which(?) no invariance constraints are imposed. This model provides a good baseline for comparing all subsequent models that impose invariance constraints hierarchically. According to the substantive interests and previous factorial invariance studies (e.g., Marsh, 1994), the hierarchical tests of equality were conducted in the following order: factor loadings for traits, factor loadings for methods, factor correlations for traits, factor correlations for methods, and residual variances.

Statistically significant change in chi-square, increment of the number of statistically significant, CFI, and AIC indicated similar patterns. That is, lack of invariance was detected in factor loadings for traits and methods, and some parts of factor correlations (methods), and, especially, residual variances (significant chi-square change, large increment of the number of significant constraints, subsequently sharp decrease in CFI, and relatively large AIC). On the other hand, invariance of factor correlations for

Table 4

Summary ofGoodness of Fit forEachGroupwithNoConstraints and All GroupswithHierarchicalConstraints.

Model	AIC	2	df	CFI	2d	dfd	ofSig.Constrainsts
NoEqualityConstraints							
Grade 5/Female	141.58	253.58	56	.905			
Grade 5/Male	107.31	219.32	56	.905			
Grade 6/Female	109.81	221.81	56	.899			
Grade 6/Male	46.69	158.69	56	.913			
Total(AcrossFourGroups)							
No Equality Constraints	405.16	853.36	224	.905			
ConstraintsFL(T)	479.56	1011.56	266	.887	158.20*	42	8
ConstraintsFL(T,M)	478.02	1100.02	311	.881	88.46*	45	9
ConstraintsFL(T,M), FC(T)	471.23	1111.23	320	.881	11.21	9	1
ConstraintsFL(T,M), FC(T,M)	472.25	1130.25	329	.879	19.02+	9	4
ConstraintsFL,FC,R	525.74	1267.74	371	.865	137.49*	42	14

Notes.FL=factorloadings, FC=factor correlation,R=Residual,T=Trait, M=Method, AIC=Akaike InformationCriterion;CFI=Comparative Fit Index; 2d anddfdindicatesubsequentdifferencein2anddffromlessconstraintstomoreconstraintsinthemodel;The 2(df=910 for the Nullmodels are 2173.50 (Grade 5/Female),1802.15(Grade 5/Male), 1753.70(Grade 6/Female),and1277.85(Grade6/Male) ; of Sig. Constraintsrefers to increment of thenumberofstatisticallysignificantconstraints(p<.05, univariateLagrangeMultipliertest),when moving Towardamorerestrictivemodel.

+ p<.05 * p<.01

traits was rather supported. Because the hierarchical tests indicated lack of invariance in the set of parameters without pinpointing the particular estimate, it was necessary to examine the source of lack of invariance in the factor structure.

In tables 5 to 8, a detailed description of the factor structure was provided with parameter estimates in the starting model (no invariance constraints). There were also tests of equality constraints in each parameter so that we could identify any lack of invariance across four groups. In Table 5, trait factor loadings were reasonable and positive. Some part of equality constraints seemed to be inappropriate in writing Quality and Writing Fluency. On the other hand, invariance of factor loadings of Adequacy of Understudying across four groups was supported. In method factor loadings, several estimates ofeach method showed lack of invariance across four groups (Table 6). In Table 7, the trait factor correlations between Writing Quality and Writing Fluency were problematic when the parameters were imposed to be invariant. As indicated above *see Table 4), there was a lack of invariance in all method factor correlations. Lastly, in Table 8, most components of the residual variances showed a lack of invariance.

D. Invariance Across Grade Within Each Gender and Across Gender Within Each Grade

As Marsh (1994) has shown the possibilities of testing the effects of gender, age, and interaction on the structure of academic self-concept, I tried to disentangle the similar effects on the MTMM structure in order to examine the

Table 5

Estimates for the Model with Three Traits and Three MethodswithNoConstraints;TraitFactorLoadings

		Factor LoadingsforTraits(Unstandardized/Standardized)			
Trait	Measured Variables	Grade 5-Female	Grade 5-Male	Grade 6-Female	Grade 6-Male
Trait 1	Understanding (Q1&Q2)	.209/.341	.195/.402	.318/.523	.218/.437
(Adequancyof	Understanding (Q3)	.722/.756	.898/.943	.523/.527	.413/.420
Understanding)					
Trait 2	Q1Claritya	.315/.392	.349/.443	.420/.579	.413/.420
(Writing	Q2Claritya	.204/.331	.277/.405	.296/.531	.262/.452
Quality)	Q3Claritya	.537/.771	.619/.880	.386/.596	.300/.444
	Q1Grammar	.141/.221	.311/.412	.339/.525	.376/.529
	Q2Grammara	.066/.119	.206/.342	.274/.494	.435/.739
	Q3Grammara	.108/.194	.272/.476	.324/.580	.409/.723
Trait 3	Q1Complexity	.385/.404	.454/.513	.512/.536	.437/.480
(Writing	Q2Complexitya	.719/.753	.523/.723	.589/.587	.344/.432
Fluency)	Q3Complexitya	.280/.299	.314/.422	.788/.793	.645/.734
	Q1No.ofWords	.335/.410	.441/.615	.491/.553	.421/.550
	Q2No.ofWordsa	.525/.637	.482/.747	.488/.578	.397/.560
	Q3No.ofWords	.211/.276	.313/.541	.759/.934	.569/.881

a When theparameters in the structure were imposed to be invariant across4groups(6constraints), at leastonesetofequalityconstraintsseemedtobe in appropriate,accordingtotheunivariateLagrangeMultipliertest ($p<.05$)

Table 6

Estimates for the Model with Three Traits and Three MethodswithNoConstraints;MethodFactorLoadings.

		Factor LoadingsforMethods(Unstandardized/Standardized)			
Method	Measured Variables	Grade 5-Female	Grade 5-Male	Grade 6-Female	Grade 6-Male
Method 1	Understanding(Q1 and Q2)a	.146/.238	.045/.093	.104/.172	.153/.307
(Question 1)	Q1Clarity	.434/.540	.453/.575	.268/.369	.297/.407
	Q1Grammar	.314/.472	.222/.293	.174/.270	.118/.166
	Q1Complexity	.752/.789	.628/.708	.764/.800	.742/.814
	Q1No.ofWordsa	.628/.770	.459/.640	.594/.669	.452/.591
Method 2	Understanding(Q1 and Q2) a	.267/.435	.322/.664	.147/.242	.054/.109
(Question 2)	Q2Claritya	.320/.518	.466/.682	.168/.301	.065/.113
	Q2Grammara	.334/.602	.198/.329	.099/.179	.019/.033
	Q2Complexity	.511/.535	.307/.424	.654/.652	.720/.902
	Q2No.ofWords	.496/.601	.311/.482	.580/.688	.392/.553
Method 3	Understanding (Q3)	.329/.344	.011/.012f	.817/.824	.812/.827
(Question 3)	Q3Clarity	.186/.268	-.057/-.080	.368/.569	.467/.691
	Q3Grammar	.222/.400	.101/.177	.074/.133	.021/.037
	Q3Complexity	.689/.735	.481/.647	.262/.264	.204/.232
	Q3No.ofWordsa	.689/.901	.453/.784	.084/.104	.111/.165

a When theparameters in the structure were imposed to be invariant across4groups(6constraints), at leastonesetofequalityconstraintsseemedtobe in appropriate,accordingtotheunivariateLagrangeMultipliertest ($p<.05$)

Table 7.
Estimates for the Model with Three Traits and Three MethodswithNoConstraints:FactorCorrelations

	Grade 5-Female	Grade 5-Male	Grade 6-Female	Grade 6-Male
CorrelationsBetweenTraits				
Understanding andWritingQuality	.999	.862	.971	.896
Understanding andWritingFluency	.459	.369	.471	.590
Writing Quality andWritingFluency a	.292	.379	.479	.522
CorrelationsBetweenMethods				
Question 1 and Question 2 a	.475	.555	.344	.165
Question 2 and Question 3 a	.312	.215	.096	.018
Question 3 and Question 3 a	.479	.326	.019	.097

a When theparameters in the structure were imposed to be invariant across4groups(6constraints), at leastonesetofequalityconstraintsseemedtobe in appropriate,accordingtotheunivariateLagrangeMultipliertest ($p<.05$)

Table 8

Estimates for the Model with Three Traits and Three MethodswithNoConstraints:ResidualVariance

	ResidualVariance(Unstandardized/Standardized)			
Measured Variables	Grade 5-Female	Grade 5-Male	Grade 6-Female	Grade 6-Male
Understanding (Q1&Q2)a	.203/.735	.075/.568	.225/.781	.173/.832
Understanding (Q3)	.283/.557	.100/.332	.043/.209	.136/.375
Q1Clarity	.357/.745	.295/.688	.279/.727	.370/.835
Q2Clarity a	.237/.789	.173/.609	.195/.792	.264/.885
Q3Clarity a	.161/.577	.108/.468	.134/.566	.148/.570
Q1Grammara	.325/.856	.425/.863	.271/.807	.350/.832
Q2Grammar	.192/.789	.281/.880	.222/.851	.157/.673
Q3Grammar	.247/.896	.242/.861	.202/.804	.153/.690
Q1Complexitya	.195/.463	.186/.486	.067/.271	.089/.327
Q2Complexitya	.134/.383	.156/.546	.231/.479	.000/.000
Q2Complexitya	.327/.609	.223/.635	.298/.549	.314/.638
Q1No.ofWordsa	.158/.488	.109/.460	.195/.497	.205/.591
Q2No.ofWordsa	.159/.483	.087/.458	.137/.439	.191/.617
Q3No.ofWordsa	.066/.335	.031/.303	.078/.343	.090/.443

a When theparameters in the structure were imposed to be invariant across4groups(6constraints), at leastonesetofequalityconstraintsseemedtobe in appropriate,accordingtotheunivariateLagrangeMultipliertest ($p<.05$)

factorial invariance as a function of gender, grade, and their joint effect. In Table 9, the first set of hierarchical models (grade 5 across gender) impose invariance over gender (boys and girls) in grade 5, and the second set of models (grade 6 across gender) impose invariance across gender in grade 6, In other words, invariance constraints over gender (boys and girls) were imposed in separate analyses of grade 5 and grade 6, and the chi-square and df from their separate analyses were summed for total models (the third set of models: across gender within grade). The results showed a similar pattern of lack of invariance (factor loadings and residual variances) in the previous four-group analyses (see Table 4).

However, for sixth graders, invariance in method factor loadings and factor correlations (traits and methods) across gender seemed to be acceptable (insignificant chi-square change, stable CFI, and smaller AIC). This six-grade-model with both factor loadings and factor correlations invariant across gender was still able to fit to the data (CFI=.90). In the total models (across gender within grade), only trait- and method factor correlations seemed to be invariant (insignificant chi-square change).

In table 10, we also imposed invariance constraints over grade levels in separate analyses of boys and girls, and then summed the chi-square and df from their separate analyses for total models (the third setofmodels:across grade within gender). For girls, invariance in method factor loadings and trait and method factor correlations could be properly imposed. In total models (across grade within gender), factor correlations (both trait and method) seemed to be invariant.

Table 9

Summary ofGoodness of Fit forInvarianceConstraintsAcrossGenderwithinGrade

Model	AIC	x2	df	CFI	2d	dfd	ofSig. Constraints
Grade5AcrossGender							
No Equality Constraints	248.97	472.87	112	0.905			
ConstraintsFL(T)	253.84	505.84	126	0.900	32.97*	14	3
ConstraintsFL(T,M)	278.25	560.25	141	0.889	54.41*	15	5
ConstraintsFL(T,M) FC(T)	277.23	565.23	144	0.889	4.98	3	0
ConstraintsFL(T,M) FC(T,M)	279.37	573.37	147	0.888	8.14+	3	2
ConstraintsFL,FC,R	334.05	656.05	161	0.870	82.68*	14	7
Grade6AcrossGender							
No Equality Constraints	156.49	380.49	112	0.905			
ConstraintsFL(T)	154.38	406.38	126	0.901	25.89+	14	1
ConstraintsFL(T,M)	141.88	423.88	141	0.900	17.50	15	1
ConstraintsFL(T,M) FC(T)	139.54	427.54	144	0.900	3.66	3	0
ConstraintsFL(T,M) FC(T,M)	135.80	429.80	147	0.900	2.26	3	0
ConstraintsFL,FC,R	131.96	453.96	161	0.897	24.16+	14	2
Total(Gender-Within-Grade)							
No Equality Constraints		853.36	224				
ConstraintsFL(T)		912.22	252		58.86*	28	
ConstraintsFL(T,M)		984.13	282		71.91*	30	
ConstraintsFL(T,M) FC(T)		992.77	288		8.64	6	
ConstraintsFL(T,M) FC(T,M)		1003.17	294		10.40	6	
ConstraintsFL,FC,R		1110.01	322		106.84*	28	

Notes. FL=factor loadings, FC=factor correlation, R=Residual, T=Trait, M=Method, AIC=Akaike Information Criterion ; CFI=Comparative Fit Index; 2d and dfd indicate subsequent difference in 2and df from less constraintstomoreconstraintsinthe model.

+ $p<.05$ * $p<.01$

Table 10

Summary ofGoodness of Fit forInvarianceConstraintsAcrossGradewithinGender

Model	AIC	2	df	CFI	2d	dfd	ofSig. Constraints
Female AcrossGrade							
No Equality Constraints	251.36	475.36	112	.903			
ConstraintsFL(T)	251.46	503.46	126	.899	28.10+	14	1
ConstraintsFL(T,M)	233.64	515.64	141	.899	12.18	15	0
ConstraintsFL(T,M) FC(T)	234.99	522.99	144	.898	7.35	3	1
ConstraintsFL(T,M) FC(T,M)	234.01	528.01	147	.898	5.02	3	1
Constraints F L, F C, R	231.93	553.93	161	.895	25.92+	14	1
Male AcrossGrade							
No Equality Constraints	154.01	378.00	112	.908			
ConstraintsFL(T)	217.73	469.73	126	.881	91.73*	14	2
ConstraintsFL(T,M)	222.73	504.73	141	.874	35.00*	15	3
ConstraintsFL(T,M) FC(T)	218.48	596.48	144	.875	1.75	3	0
ConstraintsFL(T,M) FC(T,M)	216.99	510.99	147	.874	4.51	3	0
Constraints F L, F C, R	233.97	555.97	161	.864	44.98*	14	6
Total(Grade-Within-Gender)							
No Equality Constraints		853.36	224				
ConstraintsFL(T)		973.19	252		119.83*	28	
ConstraintsFL(T,M)		1020.37	282		47.18+	30	
ConstraintsFL(T,M) FC(T)		1029.47	288		9.10	6	
ConstraintsFL(T,M) FC(T,M)		1039.00	294		9.53	6	
Constraints F L, F C, R		1109.90	322		70.90	28	

Notes. FL=factor loadings, FC=factor correlation, R=Residual, T=Trait, M=Method, AIC=Akaike Information Criterion ; CFI=Comparative Fit Index; 2d and dfd indicate subsequent difference in 2 and df from less constraintstomoreconstraintsinthe model.

+ p<.05 * p<.01

E. Summary of Effects of Grade, Gender, and Their Interaction on the MTMM Structure

The detailed analyses of various sets of hierarchical models in indicated that only some portion of the MTMM structure was invariant across gender and grade. There was also a joint effect of gender and grade on invariance of MTMM structure. To sum up, the results suggested:

1. Trait factor loadings showed a lack of invariance across gender and grade. The lack of fit was due to the inappropriateness of equality constraints across groups in the measured variables of Writing Quality and Writing Fluency.
2. Invariance of method factor loadings was influenced by joint effects of gender and grade. The invariance for sixth graders across gender, not for fifth graders, was supported. Also, the equality constraints across grade for girls seemed to be appropriate, but not for boys.
3. Factor correlations for traits seemed to be invariant across gender and grade. Yet, invariance of factor correlations for methods was weakly supported.
4. There was a lack of invariance of residual variance due to gender and grade level.

The finding of a jointeffect of gender and grade on the factorial invariance is illustrated by the summary statistics in Table 11. The first three columns in Table 11 come from the previous tables, such as total four-group (Table 4), total gender within grade (Table 9), and total grade within gender (Table 10). The chi-square and dfd values in "Gender"column are the differences between values the first column (Four

Table 11

Estimates ofGender, Grade, andInteraction EffectstotheMultitrait-multimethodStructure

Parameter Constraints		Four Groups	Gender Within-Grade	Grade Within-Gender	Gender	Grade	Interaction Equivalent
FL(T)	2d	158.20*	58.86*	119.83*	38.37*	99.34*	20.49
	dfd	42	28	28	14	14	14
FL(M)	2d	88.46*	71.91*	47.18+	41.28*	16.55	30.63*
	dfd	45	30	30	15	15	15
FC(T)	2d	11.21	8.64	9.10	2.11	2.57	6.53
	dfd	9	6	6	3	3	3
FC(M)	2d	19.02+	10.40	9.53	9.49+	8.62+	0.91
	dfd	9	6	6	3	3	3
R	2d	137.49*	106.84*	70.90*	66.59*	30.65*	40.25*
	dfd	42	28	28	14	14	14

Notes. FL=factor loadings, FC=factor correlation, R=Residual, T=Trait, M=Method, AIC=Akaike Information Criterion ; CFI=Comparative Fit Index; 2d and dfd indicate subsequentdifferencein2 anddffromlessconstraints to more constraints in the model.

+ p<.05 * p<.01

Groups) and the third column (Grade-Within-Gender). Likewise, the chi-square and dfd values in "Grade" column are the differences between values in the first column (Four Groups) and the second column (Gender-Within-Grade). Values pertinent to "interaction" were determined by subtracting values in the fourth (Gender) and fifth (Grade) columns from the first column (Four Groups). According to this overview, there were simple main effects of gender and grade in trait factor loadings and method factor correlations. A joint effect of gender and grade was found in method factor loadings and residual variances.

Ⅲ. General Discussion

This investigation examined the reliability and construct validity of a performance measure of reading comprehension and writing ability. The application of analytical scoring criteria to students' written responses to questions about their understanding of a passage of text by multiple raters yielded 14 scores that were found to be very reliable. Analysis of there scores revealed three trait factors which were significantly correlated (Writing Quality, Writing Fluency, and Adequacy of Understanding), as well as strong method (question) effects. Although significantly intercorrelated (particularly Writing Quality and Adequacy ofUnderstanding), the three traits demonstrated both convergent and discriminant validity. This three-trait three-method model was found to fit the data for boys and girls, and for fifth and sixth grade students well separately, although the factorial invariance across gender and grade was not fully supported.

Most interestingly, in the traits factors, factor correlations seemed to be stable while factor loadings showed a lack of invariance across gender, due not to Adequacy of Understanding but to the measured variables of writing components in the assessment (Writing Quality and Writing Fluency). This finding corresponded somewhat to the notion of gender stereotypic model. That is, girls perform better on constructed-response items because of some attributes in which girls are strong (i.e., writing proficiency). A detailed inspection of the estimates in the factor structure as a function of gender and grade is beyond the scope of this study and requires another systematic sample and theoretical background. It, however, would be a worthy candidate for future research.

As shown previously, scores from performance assessments using constructed response are likely to be question-specific. In many cases, such as the present instance, a simple exploratory analysis is unable to disentangle the trait and method effects, and therefore cannot adequately reveal the complex structure of the data. MTMM analysis is an effective tool for investigating the construct validity of this sort of multidimensional measure. Through CFA, MTMM analysis has some advantages over the traditional MTMM matrix using correlations, such as (a) examining the relationship between important traits, in school learning explicitly, (b) investigating the parameters as well as the measured variables, (c) evaluating alternative models in terms of constraining the relationships between variables, (d) removing method effects from the estimates of traits.

In general, every measure can be considered to be a construct-method unit (Messick, 1993). Method variance includes all systematic effects associated with particular measurement procedure that are extraneous to the focal construct being measured. The validity study, under MTMM analysis, is a systematic inquiry on construct-irrelevant variance and construct underrepresentation (Bennett, 1993; Messick, 1989). With an explicit construct network, one can differentiate the traits (construct-relevant variance) from the method effects (construct-irrelevant variance). The distinction between construct relevancy and irrelevancy is not absolute, but depends,tosomedegree, on the construct network in the particular context. The questions are considered construct-irrelevant(method) factors in the present example, but they could beconsidered part of a construct-relevant factor, if one assumed that the answer to a particular question required some unique instructionally relevant prior knowledge.

Throughout this investigation, we do recognize the exploratory nature of the analyses and also note several limitations of interpretations. First, there was a hierarchical structure in the data (Byrk & Raudenbush, 1992). Students were within the schools which belong to different districts. The multilevel covariance structure analysis cannot be implemented by the current standard programs such as LISREL or EQS so that these "design effects"were not properly specified. Second, a possibility of multiplicative models for the current MTMM structure was not explored (Cudeck, 1988), because, as asserted by Marsh (1995), we wanted to focus on the trait and method components associated with this hypothesized trait-method combination in performance assessment, and, ultimately, on the interpretation and improvement instruments.

This study is a preliminary step toward broadening and balancing the use of psychometric approaches in performance assessment. The scope of validity in any educational assessment extends to represent the meaningful construct network, and irrelevant effects are revealed more systematically. To maximize the utility of this dynamic approach to assessment, inclusive and complementary construct validation is needed. Research into ways of doing this will encompass psychometrics as well as substantial theoretical background in psychology and education.

References

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika,* 52, 317-332.

American Educational Research Association, American Psychological Asscociation, and the National Council on Measurement in Education. (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Baron, J.B. (1991). Performance assessment: Blurring the edges of assessment, curriculum, and instruction, In G. Kulm & S.M. Malcolm(Eds.), *Science assessment in the*

service of reform (pp. 247-266). Washington, DC: American Association for the Advancement of Science.

Benntt, R. E. (1993). On the meanings of Constructed response. In R.E. Bennett & W.C. Ward(Eds.), *Constructed versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 1-28). Hilsdale, New Jersey: Lawrence Erlbaum Associates.

Bentler, P. M. (1989). *EOS: Structural Equations Program Manual.* Los Angeles: BMDP Statistical Software.

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

Berk, R. A. (1986). Performance assessment: Methods and applications. Baltimore, MD: The Johns Hopkins University Press.

Bollen, K. A., & Long, J. S. (1993). Introduction. In K. A. Bollen & J. S. Long(Eds.), *Testing structural equation models* (pp. 1-9). Newbury, CA: Sage.

Born, M. P., Bleichrodt, N., & Van Der Flier, H. (1987). Cross-cultural comparison of sex-related differences on intelligence tests: A meta-analysis. *Journals of Cross-Cultural Psychology, 18,* 2823-314.

Breland, H. M., Camp, R., Jones, R. J., Morris, M. M., & Rock, D. A. (1987). *Assessing writing skill.* New York: The College Board.

Bryk, A., & Raudenbush, S. W. (1992). *Hierarchical linear models.* Newbury Park, CA: Sage.

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

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Orlando, FL: Holt, Rinehart and Winston.

Cudeck, R., (1988). Multiplicative models and MTMM matrices. *Journal of Educational Statistics, 13(2),* 131-147.

Davison, M. L. (1989). *Generalizability theory.* Paper presented at the Conference on Measurement Theories and Applications, Tainan, Republic of China

De Saint-Exupery, A. (1943). *Little prince.* San Diego, CA: HarBrace

Developmental Studies Center. (1993). *Reading Comprehension Scoring Manual.* Oakland, CA: Author.

Faingold, A. (1992). Sex differences in variability in intellectual abilities: A new look at an old controversy. *Review of Educational Research, 62(1),* 61-84.

Hyde, J. S., & Linn, M. C. (1988). Gender differences in verbal ability: A meta-analysis. *Psychological Bulletin, 104,* 53-69.

Jacklin, C. N. (1989). Female and male: Issues of gender. *American Psychologist, 44,* 127-133.

Kim, D. (1992). *Toward a performance assessment with instructional relevancy and technical adequacy: The case of Curriculum-Based Measurement.* Unpublished manuscript, University of Minnesota.

Marsh, H. W. (1989). Confirmatory factor analyses of mutitrait-multimethod data: Many problems and a few solutions. *Applied Psychological Measurement, 13(4),* 334-361.

Marsh, H. W. (1993). Multitrait-multimethod analysis: Inferring each trait-method combination with multiple indicators. *Applied Measurement in Education, 6(1),* 49-81.

Marsh, H. W. (1994). Confirmatory factor analysis models of factorial invariance: A multifaceted approach. *Structural Equation Modeling, 1(1),* 5-34.

Marsh, H. W. (1995). Latent variable models of multitrait-multimethod data. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and application* (pp. 177-198). Newbury Park, CA: Sage.

Marsh, H. W., & Richards, G. E. (1988). Tennessee Self Concept Scale: Reliability, internal structure, and construct validity. *Journal of Personality and Social Psychology, 55(4),* 612-624.

Martin, J. D., & Hoover, H. D. (1987). Sex differences in educational achievement: A longitudinal study. *Journals of Early Adolescence, 7,* 65-83.

Mehrens, W. (1992). Using performance assessment for accountability purposes. *Educational Measurement: Issues and Practice, 11(1),* 3-9 and 20.

Messick, S. (1993). Trait equivalence as construct validity. In R. E. Bennett & W. C. Ward (Eds.), *Constructed versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 61-74). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

National Assessment of Educational Progress. (1984). Reading, thinking and writing: Results from the 1979-80 national assessment of reading and literature (Report No. 11-L-01). Princeton: ETS.

Priestley, M. (1982) *Performance assessment in education and training: Alternative techniques.* Englewood Cliffs, NJ: Educational Technology Publications.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer.* Newburry Park, CA: Sage.

Stiggins, R. J., & Bridgeford, N. J. (1984). *Performance assessment for teacher development.* Northwest Regional Educational Lab, Portland, OR: Center for Performance Assessment.

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

Wothke, W., & Browne, M. (1990). The direct product model for the MTMM matrix parameterized as a second order factor analysis model. *Psychometrika, 55(2),* 255-262.