DOCUMENT RESUME

ED 325 506 TM 015 739

AUTHOR Tindal, Gerald; Nolet, Victor

TITLE The Construct Validity of Curriculum-Based Measures of Achievement: A Multitrait-Multimethod Analysis.

PUB DATE Apr 90

NOTE 29p.; Paper presented at the Annual Meeting of the American Educational Research Association (Boston, MA, April 16-20, 1990).

AVAILABLE FROM 275 Teacher Education, University of Oregon, Eugene, OR 97405-1213.

PUB TYPE Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.

DESCRIPTORS *Achievement Tests; *Construct Validity; Elementary Education; *Elementary School Students; Language Skills; Mathematics Achievement; *Multitrait Multimethod Techniques; Reading Achievement; *Skill Analysis; Spelling; *Test Validity; Writing Achievement

IDENTIFIERS Convergent Validation; *Curriculum Based Assessment; Divergent Validation; Stanford Achievement Tests

ABSTRACT

Convergent validity and divergent validity were studied in several basic skill areas using the multitrait-multimethod procedure outlined by D. T. Campbell and D. W. Fiske (1959). The academic skill domains considered were: (1) reading; (2) mathematics; (3) writing and language arts; and (4) spelling. A total of 297 students in grades 1 through 5 served as subjects (65 in grade 1, 57 in grade 2, 59 in grade 3, 60 in grade 4, and 56 in grade 5). The multitrait-multimethod matrix was established using the Stanford Achievement Test (SAT) and some curriculum-based measures. The results do not support specific skill independence, even across such major domains as reading and mathematics using very different methods to measure proficiency. Although convergent validity was found, divergent validity was almost entirely lacking. The findings support a uniform conception of achievement that is globally defined in terms of components and methods. Six tables contain data from the study. (SLD)

The Construct Validity of Curriculum-based Measures of Achievement: A Multitrait-multimethod Analysis

Gerald Tindal

Victor Nolet

University of Oregon

Running Head: Multitrait-multimethod Analysis

Paper presented at the American Educational Research Association, Boston, MA, April 16-20, 1990.

Requests for reprints: Dr. Gerald Tindal, 275 Teacher Education, University of Oregon, Eugene, OR. 97405-1213

Abstract

The construct validity of many psychological traits and concepts (e.g. intelligence, giftedness, learning disabilities, etc.) have been addressed over the past thirty years; however, the construct of achievement has not been adequately considered. At best, limited attention has been devoted to various constructs within specific skill areas, such as the construct of comprehension in reading, , or the construct of problem-solving in math. This study focused on construct validity across several basic academic skills domains: reading, math, writing-language arts, and spelling. A multitrait-multimethod analysis was conducted to investigate convergent and divergent validity of these skills in six different grade levels. The results did not support specific skill independence, even across such major domains as reading and math, using very different methodologies to measure proficiency. Although convergent validity was found, divergent validity was almost entirely lacking. These findings support a uniform conception of achievement that is globally defined both in terms of components and methods.

The Construct Validity of Curriculum-based Measures of Achievement: A Multitrait-multimethod Analysis

Analysis of subskills in reading has been conducted intermittently for the past 20 years. The controversy generally is subsumed within a larger arena of debate on the construct validity of reading, in general, and reading comprehension, in particular. Generally, the researchers have taken a reading test and analyzed the interrelationships between various subtests. The conclusions have been fairly uniform, as indicated by three quotes noted below:

In an analysis of the Iowa Test of Basic Skills, Level 11 (Hieronymous, Lindquist, & Hoover, 1978), Cummings (1982) found that "in general, it appears that the subskill item classifications in reading comprehension used in the testing or test-development process are limited in their usefulness as a basis for diagnosing specific skill strengths or weaknesses" (p. 65).

Drahozal and Hanna (1978) analyzed the Nelson Reading Skills Test (Hanna, Schell, & Schreiner, 1977) and concluded that "the three kinds of tasks (*Literal, Translation*, and *Higher Level*) identified for this particular reading comprehension test do not reflect corresponding independent attributes at the grade levels tested (grades three through nine). Rather, the tasks seem to be alternative means of measuring the same thing" (p. 419).

Finally, in a study using three different published reading tests (McCullough Word Analysis Tests, 1963; Gates-MacGinitie Reading Test, Survey D, 1965; and California Reading Test, Elementary by Tiegs & Clark, 1963), and employing convergent -divergent validity methodology, Farr and Roelke (1971) found that "in most cases the correlations of different skills with each other when measured by the same method were higher than the same skills when measured by different methods" (p. 31). The study was an extension of an earlier investigation conducted by Farr (1968) that had arrived at similar conclusions.

Although the earlier studies had investigated the degree to which two subtests are interrelated, the focus has been only on singular methodologies. This last study is potentially the most interesting and complete. By using convergent-divergent validity, Farr (1968) and Farr and Roelke (1971) considered a broader and more comprehensive question, looking at different skills measured in different ways. They conceptualized three traits as being measured: word attack, vocabulary, and comprehension. Furthermore, three methods for measuring these traits were included, two of which employed teacher ratings and one of which was based on published test scores.

Although they describe their study as one in which multitrait-multimethod construct validity was being used (Campbell & Fiske, 1959), a singular trait was actually being studied: reading. Furthermore, only two, not three methodologies were employed, since two of the methods were ratings that differed only in terms of the *source* of measurement, rather than the *method* of measurement. Further, ratings of reading skill were not really direct assessments of reading at all, but perceptions based on an unknown amalgam of information which teachers favor to varying degrees (Salmon-Cox, 1981). Finally, only one age group of students (fifth graders) was considered, precluding any analysis of developmental trends in the relationships among the measures. Reading theorists currently disagree whether reading should be viewed as a sequential series of independent skills that emerge over time or as a holistic skill which cannot be subdivided but develops somewhat uniformly over time (Farr & Carey, 1986). This issue may be addressed, in part, by looking at the independence of subskills over the time period in which reading proficiency develops.

To overcome these limitations, a similar study was conducted with the following modifications. First, a true multitrait-multimethod matrix was constructed: The traits included other basic skill areas (spelling, math, and language arts), in addition to reading. Method variance included two direct measures of basic skills, rather than one direct and one indirect method (perceptions of reading). Finally, students from several grade levels were included, to assay any developmental trends in the existence of subskillindependence.

In this study, the multitrait-multimethod matrix was established with the Stanford Achievement Test (SAT) (Gardner, Rudman, Karlsen, & Mervin, 1982) and curriculum-based measures (CBM) as developed by Deno and associates (Deno, Mirkin, Chiang, & Lowry, 1980; Deno, Mirkin, & Marston, 1980; Deno, Mirkin, Lowry, & Kuehnle, 1980).

The SAT has several skill areas that generally are comprised of two subtests each, including the following: reading, listening, spelling/language, and math. On all subtests, the student selects the correct answer from an array of alternate choices. Curriculum-based measures, in contrast employ production responses, in which the student creates the answer. In reading, a count is made of the number of words read correctly in one minute from a basal passage. In spelling, the measure is comprised of a rolling dictation of vocabulary words, with the student writing them down individually on lined paper for 2 minutes; a count is made of the number of words spelled correctly and the number of correct letter sequences. The writing measure uses a story starter to generate a 3-minute sample of creative writing, which is then scored for the number of total words, correctly spelled words, and correct word sequences. Finally, in math, numerous single- or multiple-operation computation problems are presented to the student within a 2-minute period and a count is taken of the number of digits in the correct places. See Shinn, (1989) for a complete description of the research and uses to which these basic skills measures have been applied.

Both measurement systems, the published achievement test and the CBM, have extensive technical adequacy data documented either in technical manuals or the professional literature. A considerable research base on the concurrent validity of the CBM procedures has employed published, norm-referenced tests as the criterion. However, single-skill subtests have typically been administered, rather than complete batteries. In this study, complete test batteries were administered.

Method

The study was conducted in a small northwest community located in a moderate sized urban area of 150,000. All students came from one elementary school within this community. A total of 297 students in grades 1 through 5 served as subjects, including 65 from Grade 1, 57 from Grade 2, 59 from Grade 3, 60 from Grade 4, and 56 from Grade 5. Their achievement levels (using standard scores) on major subtest skill areas of the SAT (Gardner et al., 1982) are summarized in Table 1.

Insert Table1 about here

All measures were administered at the end of the school year using standardized procedures. The SAT was administered in April as part of the normal district testing routine, with four different measures included: reading, listening, spelling/language, and math. Individual classroom teachers administered the SAT subtests over a one-week period. The CBM measures were administered one month later as part of the school's development of normative performance levels to be used as screening criteria in special education. All curriculum-based measures were administered by trained Master's and doctoral students from a nearby university; only the reading measure was individually administered, with spelling, math, and writing administered to intact classrooms. Five different CBM skill areas were included: reading, spelling, writing, math, and tool movements. The last measure was included for Grades 1 through 3 to determine the relative impact of proficiency in pencil manipulation on spelling, writing, and math tasks. As described earlier, these measures are all production responses of brief duration, ranging from 1 to 3 minutes each.

The correlation matrices presented in Tables 2-6 were developed in the following manner. For CBM, the median correlation was selected from several different forms (reading and math) or from scoring the same protocol in several different ways (spelling and writing). For the SAT, the median correlation was selected from several subtests with each other and with the total skill area score.

Insert Tables 2 through 6 about here

To interpret the data, the following suggestions are offered. First, reliability coefficients, reflecting intercorrelations between various measures of the same trait measured in the same manner, should be the highest in the tables. Second, validity coefficients, representing the same trait measured in different ways, should all be greater than zero and large. Third, these validity coefficients should be greater than all other coefficients in the same rows and columns within the heterotrait-heteromethod area of the correlation matrix. Fourth, the validity coefficients should be larger than any coefficients in the heterotrait-monomethod area of the correlation matrix. Finally, the pattern of correlations should be the same for the traits in both the heteromethod and monomethod areas of the matrix.

Results

Subject achievement data are displayed in Table 1. The student population is slightly above average in most skill areas and in most grade levels. The multitrait-multimethod data have been summarized by grade level in Tables 2 through 6, with each correlation matrix displaying the relationships between these basic skills for both methods.

In all grades, the reliability coefficients are the highest and exceed all other coefficients. The next highest coefficients tend to be the validity coefficients, with the exception of math, which is very modest in grades 2-4; other exceptions are noted below. All coefficients are significantly greater than zero.

Most of these validity coefficients exceed the heterotrait-heteromethod coefficients, with the following two exceptions. First, because the math validity coefficients are low in three grades (2-4), they are not sufficiently large to exceed the relationships between other traits measured in other ways. Further, the relationship between math and reading, spelling, and writing is quite strong in four grades (2-5). Second, in virtually all grades, the relationship between reading and spelling is as strong as that between the same measures using different methods (i.e., reading from the SAT and CBM, or spelling from the SAT and CBM).

The validity coefficients do not consistently exceed different traits measured in the same manner. Because the relationship between other traits (particularly spelling) and reading measures is so strong, regardless of method, the validity coefficients rarely exceed it. The intercorrelations between the different traits measured with the SAT are also typically high, across all grade levels.

Finally, the pattern of correlations is somewhat consistent across both the hetero- and mono-methods. The highest relationships tend to occur with reading and spelling, then reading and math, and finally, between spelling and math.

Discussion

This study investigated convergent and divergent validity in several basic skill areas using the multitrait-multimethod procedure outlined by Campbell and Fiske (1959). The results are similar to other studies conducted within the area of reading comprehension.

The most consistent finding was that all diagonal coefficients within methods were the highest. However, in constructing the correlation tables, the manner in which these coefficients were selected may have influenced this outcome. For example, the median coefficients for all separate subtests within and across each measurement methodology were selected. With the reading CBM, four different measures were administered, two grade level passages, one passage from the lowest reading group in each grade level, and a curriculum-free word list. In math, three different probes were administered (varying in the operations sampled), and the median correlation selected. Both spelling and writing measures employed correlations between different scoring systems from the same protocol. For the SAT, three coefficients typically were used, reflecting the correlation between two subtests and their correlation with the skill area subtotal. This strategy may actually inflate the coefficients, since the response format was exactly the same and only the item sampling procedures differed. Furthermore, with the SAT, each subtest is inherently included within the skill area subtotal. With the CBM areas using the same protocol (i.e., spelling and writing), an inherent dependency may exist between the different scoring systems. Nevertheless, as these coefficients reflect the reliability of the monomethod measures, these artifacts are in keeping with the logic behind the matrix.

Probably the most significant dimension which distinguished the two methods from each other was the type of response. A production response was employed with curriculum based measures, while the SAT, like most group administered published, norm-referenced achievement tests, used a selection response. Each type of response has an artifact that may influence performance and subsequently affect the method variance.

With three of the CBM tasks—writing (in all grades except First), spelling, and math—minimal proficiency in basic tool movements (White and Haring. 1980), is an issue. Certainly in Grades 1 and 2, performance on these tasks may be influenced by skill in manipulating a pencil, which is unrelated to proficiency in the academic skill. Because we were concerned with this issue at the beginning of the study, we gave students a series of tasks primarily free of academic components or cognitive demands that consisted primarily of rote tasks reflecting skill in moving a pencil. These tasks included printing or writing their name, writing the numbers 1 through 9, and writing the letters a-z. All tasks were timed for a brief period (30 seconds to 1 minute), with students repeatedly performing each task during this interval. Each test score was then computed as the number of letters or numbers completed correctly. The relationship between these non-academic tasks and other academic ones was either zero or moderately negative. Although the CBM tasks rely upon fine motor skills, this component does not influence the rank ordering of students on academic skills of writing, spelling, or math.

In contrast, the SAT uses multiple-choice responses, which may present problems with dependencies (context cues) across alternative items and between the information in the passage and the test items. For example, Tuinman (1974) investigated five reading comprehension tests and found that test takers could answer many items without having read the passage. Furthermore, if students know the answer for a specific item, they may have a greater probability of answering other items correctly. And, as Johnston (1984) found, student prior knowledge is a major source of bias in many published achievement tests.

Although the two measurement systems employed considerably different responses, the monotrait-heteromethod correlations were quite strong. These coefficients, representing validities in the matrix, should have been higher than either of the two types of heterotrait correlations, whether mono- or hetero-method. The finding of strong correlations among different measures of reading, spelling, and written expression is not new; many other studies have been completed with very consistent findings (Marston, 1989). These findings, however, have been confined to only one trait. A clear exception was the weak relationship between the two different math subtests. Because this coefficient was so low, the relationship between math and reading or spelling exceeded it, whether using one or both methods to assess it.

The findings from this study are quite similar to those reported by Farr and Roelke (1971). While they found convergent validity, they also found "an almost total lack of discriminant validity" (p 32). We also found convergent validity with our measures. And, like Farr and Roelke, we found serious problems with the discriminant or divergent validity of the traits and methods. Different traits should not have correlated with each other as well as the same trait correlated across different methods for measuring it.

A finding that was somewhat surprising was the correlation between reading and spelling, regardless of the manner in which it was measured. And the low correlation between the math measures, whether tested with similar or different methods, resulted in similar if not higher correlations between math and all the other skill traits. Finally, different skills within the SAT were all quite highly intercorrelated.

Farr and Roelke (1971) explained their lack of discriminant validity as a function of either invalid instruments or inseparable subskills. If we assume that, because of the many criterion validity studies that have been conducted with both the CBM and SAT, the measures we used are valid, we must conclude that the subskills may not be separate. However, unlike the Farr and Roelke study, we investigated very different skill areas, not just subsets within a skill area (like reading comprehension). They suggest that the method variance is responsible for this outcome; yet, we assert that the trait separation is suspect. Although we have traditionally separated achievement tests into separate skill areas, the high correlations among them suggest that they may be part of a larger response class reflecting generalized achievement. This global construct simply reflects overall achievement. Further subdivisions into separate traits we consider ill-advised. This argument is further strengthened by noting the size of these relationships across measurement methodologies. How else can the high relationship be explained between spelling words from dictation or solving math problems and silently reading passages and then answering comprehension comprehension questions?

References

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

Cummings, O. W. (1982). Differential measurement of reading comprehension skills for students with discrepant subskill profiles. Journal of Educational Measurement, 19, 59-66

Deno, S. L., Mirkin, P. K., & Marston, D. (1980). Relationships among simple measures of written expression and performance on standardized achievement tests. (Research Report No. 22). Minneapolis, MN: University of Minnesota Institute for Research on Learning Disabilities.

Deno, S. L., Mirkin, P. K., Chiang, B., & Lowry, L. (1980). Relationships among simple measures of reading and performance on standardized achievement tests. (Research Report No. 20). Minneapolis: University of Minnesota Institute for Research on Learning Disabilities.

Deno, S. L., Mirkin, P. K., Lowry, L., & Kuehnle, K. (1980). Relationships among simple measures of spelling and performance on standardized achievement tests. (Research Report No. 21). Minneapolis, MN: University of Minnesota Institute for Research on Learning Disabilities.

Drahozal, E. C., & Hanna, G. S. (1978). Reading comprehension subscores: Pretty bottles for ordinary wine. Journal of Reading, 21, 416-420.

Farr, R. (1968). The convergent and discriminant validity of several reading tests. In G. B. Schick & M. M. May (Eds.) Multidisciplinary aspects of college-adult reading. Yearbook of the National Reading Conference.

Farr, R., & Roelke, P. (1971). Measuring subskills of reading: between standardized tests, teachers' ratings, and reading specialists ratings. Journal of Educational Measurement, 18, 27-32.

Farr, R., & Carey, R. F. (1986). Reading: What can be measured? Newark, DE: International Reading Association.

Gardner, E. F., Rudman, H. C., Karlsen, B., & Merwin, J. C. (1982). Stanford Achievement Test-7th edition. San Antonio, TX: The Psychological Corporation.

Gates, A. I., & MacGinitie, W. H. (1978). Gates-MacGinitie reading tests. New York: Teachers College Press.

Hanna, G. S., Schell, L. M., & Schreiner, R. (1977). The Nelson Reading Skills Test. Boston, MA: Houghton Mifflin.

Hieronymous, A. N., Lindquist, E. F., & Hoover, H. D. (1978). Iowa Test of Basic Skills, Forms 7 and 8. Lombard, IL: Riverside Publishing Company.

Johnston, P. (1984). Prior knowledge and reading comprehension test bias. Reading Research Quarterly, 19, 219-239.

Marston, D. (1989). A curriculum-based measurement approach to assessing academic performance: What it is and why do it. In M. R. Shinn (Ed.) Curriculum-based measurement: Assessing special children (pp. 18-78). New York: The Guilford Press.

McCoullough, R. T. (1963). McCoullough word analysis tests. Boston: Ginn.

Salmon-Cox, L. (1981). Teachers and standardized achievement tests· What's really happening? Phi Delta Kappan, 62(May), 631-634.Tuinman (1969).

Shinn, M. R. (1989). Curriculum-based measurement: Assessing special children. New York: The Guilford Press.

Tiegs, E. W., & Clark, W. W. (1963). California reading tests. Monterey, CA: California Test Bureau.

Tuinman, J. J. (1974). Dete;mining the passage dependence of comprehension questions in five major tests. Reading Research Quarterly, 9, 206-223.

White, O., & Haring, N. (1980). Exceptional Teaching. Columbus, OH: Charles Merrill.

Table 1

Achievement levels on Stanford Achievement Test (1982) Skill Areas

Grade Level		Reading		Listening		Spelling/ Language		Mathematics	
	N	M	SD	M	SD	M	SD	M	SD
One	62-65	519	57	578	28	479	58	525	43
Two	55-57	604	47	614	29	575	48	585	35
Three	55-59	630	39	632	29	632	30	625	34
Four	56-60	645	33	651	32	640	32	636	35
Five	52-56	638	35	657	24	643	33	635	32

Table 2

Correlation Matrix for Grade 1 Displaying Relationships between Performance on Different Traits as Measured using Different Methods

	CBM-Rdg	CBM-Splg	CBM-Mth	CBM-Tool	SAT-Rdg	SAT-List	SAT-Splg	SAT-Mth
CBM-Rdg	.90							
CBM-Splg	.72	.95						
CBM-Mth	.27	.44	-					
CBM-Tool	-.05	-.03	-.10					
SAT-Rdg	.76	.84	.41	-.13	.90			
SAT-List	.39	.36	.38	-.06	.46	.77.		
SAT-Splg	.88	.76	.30	-.09	.83	.35	-	
SAT-Mth	.45	.57	.63	-.06	.61	.65	.51	.92

Table 3

Correlation Matrix for Grade 2 Displaying Relationships between Performance on Different Traits as Measured using Different Methods

	CBM-Rdg	CBM-Splg	CBM-Wrtg	CBM-Mth	CBM-Tool	SAT-Rdg	SAT-List	SAT-Splg	SAT-Mth
CBM-Rdg	.91								
CBM-Splg	.55	.93							
CBM-Wrtg	.34	.29	.88						
CBM-Mth	.40	.36	.42	.70					
CBM-Tool	-.02	-.13	.02	.17	.33				
SAT-Rdg	.59	.65	.21	.29	-.22	.85			
SAT-List	.37	.20	.07	.20	-.12	.48	.84		
SAT-Splg	.49	.58	.28	.27	-.16	.53	.20		
SAT-Mth	.39	.39	.14	.35	.12	.53	.57	.34	.80

Table 4

Correlation Matrix for Grade 3 Displaying Relationships between Performance on Different Traits as Measured using Differen

Methods

	CBM-Rdg	CBM-Splg	CBM-Wrtg	CBM-Mth	CBM-Tool	SAT-Rdg	SAT-List	SAT-Splg	SAT-Mth
CBM-Rdg	.91								
CBM-Splg	.69	.92							
CBM-Wrtg	.70	.81	.98						
CBM-Mth	.66	.61	.55	.70					
CBM-Tool	-.42	.18	-.30	.06	.51				
SAT-Rdg	.62	55	.56	.32	-.38	.90			
SAT-List	64	42	.61	.57	-.15	.72	.95		
SAT-Splg	.77	61	66	.37	-.38	.65	.63	.76	
SAT-Mth	.68	46	.56	.44	-.19	.73	.78	.61	.79

Table 5

Correlation Matrix for Grade 4 Displaying Relationships between Performance on Different Traits as Measured using Different Methods

	CBM-Rdg	CBM-Splg	CBM-Wrtg	CBM-Mth	SAT-Rdg	SAT-List	SAT-Splg	SAT-Mth
CBM-Rdg	.85							
CBM-Splg	.56	92						
CBM-Wrtg	50	43	93					
CBM-Mth	.37	.47	.31	.49				
SAT-Rdg	70	61	.47	.33	.63			
SAT-List	.45	32	.10	.19	.52	.89		
SAT-Splg	.62	.69	.46	.42	.72	.49	.87	
SAT-Mth	.37	56	24	.57	.59	.40	.55	73

Table 6

Correlation Matrix for Grade 5 Displaying Relationships between Performance on Different Traits as Measured using Different Methods

	CBM-Rdg	CBM-Splg	CBM-Wrtg	CBM-Mth	SAT-Rdg	SAT-List	SAT-Splg	SAT-Mth
CBM-Rdg	.86							
CBM-Splg	69	93						
CBM-Wrtg	53	51	90					
CBM-Mth	.48	.48	31	.64				
SAT-Rdg	66	69	27	59	.90			
SAT-List	.34	.37	-.01	.46	.64	.87		
SAT-Splg	.61	.64	.26	.61	.77	.67	.92	
SAT-Mth	.47	.49	.16	.62	.66	.68	.69	.85