Validity of the Texas Assessment of Knowledge and Skills (TAKS)

Jill Burk
Dale Johnson
&
Janet Whitley

Tarleton State University

Abstract
The study addressed the statistical validity of the Texas Assessment of Knowledge and Skills (TAKS) in reading and mathematics for 3rd and 4th grade students (n = 202). Because of the importance of decisions based on high-stakes testing results, technical qualities need to conform to high psychometric standards, including indicators of statistical validity. The case for content (objective/standard-based) validity is provided by the Texas Education Agency; however, evidence of statistical validity is lacking. The study utilized correlational methods to investigate predictive, concurrent, discriminant, convergent and construct validity of the state-mandated exam. Results confirm that the TAKS is a reasonably valid indicator of third and fourth grade reading and mathematics achievement. Validity coefficients and patterns of correlations support the hypothesis that the TAKS does measure what it purports to measure and is relatively free of influences extraneous to the mathematics and reading domains.

Introduction
The purpose of this study was to investigate the statistical validity of the Texas Assessment of Knowledge and Skills (TAKS) test, a state-mandated high stakes test for a population of elementary school students. Because of the importance of the decisions which are influenced by scores on state-mandated tests, the issues involving technical qualities of such tests are critical. For example, the Standards for Educational and Psychological Testing (1999) state:

The higher the stakes associated with a given test use, the more important it is that test-based inferences are supported with strong evidence of technical quality. In particular, when the stakes for an individual are high, the important decisions depend substantially on test performance, the test needs to exhibit higher standards of technical quality for its avowed purposes than might be expected of tests used for lower-stakes purposes. (p. 139)

The national-level No Child Left Behind (NCLB) and statewide emphasis by many states on utilizing testing results for important decisions and as a major source of evaluative
efforts leaves no doubt that state-mandated standards-based testing qualifies as high stakes testing, and as such, should itself be subjected to rigorous examination for technical qualities.

As an achievement test, presentation of evidence of various types of psychometric properties such as reliability and validity is standard testing practice. For purposes of achievement testing instrumentation, any elementary measurement text clearly spells out the various types of reliability estimates and the various types of validity evidence including content, construct, and statistical validity (e.g., Linn & Gronlund, 2000).

The Texas Education Agency (2003) provides fairly extensive descriptions of content validity procedures and internal-consistency reliability of the TAKS. Reliability coefficients in the 0.80’s coupled with lengthy descriptions of efforts to align subject-area content with the testing instrumentation provide important support for accuracy and content validity (relevance). However, other reliability estimates (stability) and statistical validation are lacking.

The Standards for Educational and Psychological Testing (1999) manual defines convergent, discriminant, and test-criterion (concurrent and predictive) validity (p. 14) as well as content validity (p. 11) and describes the importance of each (pp. 9 – 24). Information regarding the various types of statistical validity is not readily apparent from the state website nor from literature in educational and psychological data bases. One 1999 study correlating algebra I end-of-course test results with algebra numerical grades is provided by the Texas Education Agency (TEA). However, the data are not recent nor does the study address the validity of the TAKS – the Texas exam. In addressing the need for additional evidence for validity at the elementary school level, the objectives of the present study were to examine several types of validity of the high-stakes testing instrument for 3rd and 4th grade students.

The present study does not address the content validity of TAKS since ample evidence is available from the TEA website on the procedures used to develop the tests. On the other hand, various types of statistical validity were examined including: concurrent, predictive, and construct (including convergent and divergent). Concurrent and predictive validity were quantified with simple linear bivariate associations between the testing scores and performance on appropriate criteria. These were displayed as product-moment correlation coefficients (r’s) which are common in testing literature and technical reports. Construct validity was generally a more complex set of procedures, but many times utilized an analysis of patterns of correlation coefficients in a multitrait-multimethod matrix (MTMM) (Campbell and Fiske, 1959) or other theoretically based reasoning.

Related Literature

Patterns of correlations suggested by Campbell and Fiske (1959) for analyzing convergent and divergent validity utilize a procedure called the analysis of the multitrait-multimethod matrix (MTMM). Sawilowsky (2002) noted that “the most frequently cited article of that half century was by Campbell and Fiske, with over 2,000 citations. (The second most cited article, with over 900 citations, was also on construct validity.)” (p. 78). Thus, the use of the MTMM is widespread and has been used extensively in educational settings. For
example, Marsh (1992), using MTMM, found components of academic self-concept across eight subject areas for high school boys was well differentiated and surprisingly content-specific. Although complex analyses utilizing latent variable approaches (Marsh, 1993) have been found to be effective, the correlations among measured variables remains a popular and defensible procedure for school-related assessments.

MTMM methodology has also been used in lower grade settings (Hooper, 1988), 7th grade students (Marshall, 1991), and 14 year olds (Massey, 1977). Nolet and Tindal (1990) found in using the MTMM procedure that academic domains of reading, mathematics, writing & language arts, and spelling for elementary students lacked discriminant validity although convergent validity was satisfactory. They concluded that elementary level achievement is globally defined with little specific skill independence.

In contrast, Cole, Gondoli, and Peeke (1998) studied five dimensions of competence for third and sixth grade students using a rating scale with teachers and parents. They found a high level of discriminant validity. Assessments in reading, language arts, and mathematics of third and fourth graders by Crehan (2001) found little evidence for convergent or discriminant validity on a locally developed performance measure. The blame was attributed to low reliability on the performance measure due to excessive subjectivity in scoring. Although there is not total agreement regarding expectations on discriminant validity at the elementary school level, the reliability of the TAKS as an objectively scored measure was judged sufficient to use MTMM methodology.

Amrein and Berliner (2002) using ACT, SAT, NAEP and AP test data, argue that validity of high-stakes testing nationwide is weak and relate the “questionable validity” to failed policy. In contrast, Green, Winters, and Forster (2003) posit that the Amrien and Berliner data are irrelevant and provide evidence of their own from 5,587 schools that show strong concurrent validity of high-stakes testing with correlations ranging from 0.35 to 0.96 with an average correlation of 0.88. In his Presidential Address given at the Annual Meeting of the National Council on Measurement in Education, Haertel (1999) discussed the validity of high-stakes testing and how it could be improved.

Methods

Participants

Students participating in the study were 4th grade students (n = 202) attending a third and fourth grade elementary school of about 500 FTE in a central Texas community. The student body was approximately 79% white, 19% Hispanic, and 1% African American. Just over 41% were economically disadvantaged and 7% were classified as Limited English Proficient. Over 90% of the students satisfied the reading and mathematics standards on the state-mandated TAKS test. The school has a heterogeneous faculty in terms of experience ranging from one year to over 20 years – the average number of years of teaching experience for the faculty was about 16. The school is considered typical for middle-sized, non-urban schools.
Procedures

Subsequent to securing permission from the school system to use student data to conduct the study, collection of data from student records was initiated. Data included 3rd and 4th grade reading scale scores and mathematics scale scores on the TAKS test; 3rd grade reading comprehension and mathematics total scores from the Iowa Test of Basic Skills (ITBS); 3rd and 4th grade benchmark test scores (Renaissance Learning, 2004); and 3rd and 4th grade teacher reported end-of-the-year numerical grades in reading and mathematics. Then the data were analyzed and the results were interpreted to formulate conclusions.

Analyses

The study included correlational applications and analyses of patterns of associations among several pairs of variables. For assessing convergent, discriminant, and predictive validity of the reading and math scales, a modified MTMM matrix (Figure 1) was utilized. The elements in the MTMM are correlation coefficients among 3rd and 4th grade TAKS reading scale scores, 3rd and 4th grade TAKS mathematics scale scores, numerical mathematics grades and numerical reading grades for both grade levels. As shown in

Table 1

<table>
<thead>
<tr>
<th>Multitrait, Multimethod Matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grade 3 TAKS Scores</td>
</tr>
<tr>
<td>----------------------</td>
</tr>
<tr>
<td><strong>Reading</strong></td>
</tr>
<tr>
<td>Grade 3 TAKS</td>
</tr>
<tr>
<td>Reading a</td>
</tr>
<tr>
<td>Grade 4 TAKS</td>
</tr>
<tr>
<td>Reading c</td>
</tr>
<tr>
<td>Grade 4 Grades</td>
</tr>
<tr>
<td>Reading c</td>
</tr>
<tr>
<td>Grade 4 Benchmarks</td>
</tr>
</tbody>
</table>

Note: a = monotrait, monomethod; b = heterotrait, monomethod; c = monotrait, heteromethod; d = heterotrait, heteromethod

As shown in Table 1, the MTMM consisted of two traits (mathematics and reading) and three methods of assessment (TAKS test scores, teacher grades, and benchmark scores) for two consecutive years.

The coefficients labeled “a” are monotrait, monomethod coefficients because the same trait is assessed with the same method. In an MTMM analysis, the
“a” correlations should be the largest in the matrix as they represent reliability. The “c” coefficients are montrait, heteromethod associations and, as indicators of convergent validity, should represent the second largest correlations in the matrix; i.e., the various methods of assessment are said to “converge” on the traits. The “b’s” represent multitrait, monomethod associations – different traits assessed by the same method. Finally, the “d’s” are heterotrait, heteromethod coefficients and should represent the smallest values in the MTMM. The difference between the “b’s” and the “d’s” is known as method variance. The extent to which method variance is minimized reflects what is called discriminant validity (Thorndike, 1977). In order to average the associations to compare patterns, the data were first converted to Fischer’s z-score transformations for arithmetic calculations.

Concurrent validity was assessed using Pearson correlation coefficients, some of which are redundant to elements of the MTMM matrix, between the following pairs of variables:

TAKS Reading with Teacher Grades in Reading at Grade 3;
TAKS Mathematics with Teacher Grades in Mathematics at Grade 3;
TAKS Reading with Teacher Grades in Reading at Grade 4
TAKS Mathematics with Teacher Grades in Mathematics at Grade 4;
TAKS Mathematics with ITBS Mathematics Total;
TAKS Reading with ITBS Reading Comprehension;
TAKS Mathematics with Benchmark scores at Grade 3; and
TAKS Reading with Benchmark scores at Grade 3.

Predictive validity of the TAKS was estimated using Pearson correlation coefficients between the following variables:

TAKS Reading at Grade 3 with TAKS Reading at Grade 4;
TAKS Reading at Grade 3 with Teacher Grades at Grade 4;
TAKS Mathematics at Grade 3 with TAKS Mathematics at Grade 4; and
TAKS Mathematics at Grade 3 with Teacher Grades in Mathematics at Grade 4.

Finally, using a rationale that foundational underpinnings of assessment instruments coupled with time intervals of administration should influence relationships among the test scores, conjectural hypotheses relevant to construct validity were tested. More specifically, assessments which are based on standards should correlate higher with other measures based on the same standards than with measures based on content. Also, assessments at the same point in time (i.e. beginning of the year, mid-term, and end of the year) should relate higher to other assessments than for equivalent assessments executed at different times.

Results

Convergent and Divergent Validity

Results are presented in the form of a MTMM matrix followed by concurrent, predictive, and construct validity coefficients. As previously described, the MTMM matrix methodology provides predicted patterns of correlations among specified variables in the study as shown in Table 2.
As may be seen in Table 2, the diagonal (a’s) elements in the MTMM are generally the largest correlations as expected. The average (mean) correlation coefficient for the diagonal is $r = 0.77$ which can be viewed as a modified equivalent form reliability estimate. The heterotrait, heteromethod elements (d’s) are the smallest relationships having an average correlation of 0.62. The monotrait, heteromethod elements (c’s) represent convergent validity and have an average of $r = 0.70$. As previously noted, the quantity represented by the difference between the “b” and “d” elements is called “method variance.” Since the mean value of the “b” elements is $r = 0.66$ and the mean value of the “d” elements is 0.62, the method variance metric is 0.04. This value is quite small; the method variance is therefore minimal which implies a reasonable degree of discriminant validity.

**Table 2**

Multitrait, Multimethod Matrix Results

<table>
<thead>
<tr>
<th></th>
<th>Grade 3 TAKS Scores</th>
<th>Grade 3 Class Grades</th>
<th>January Renaissance Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Reading</td>
<td>Math</td>
<td>Reading</td>
</tr>
<tr>
<td>Grade 4 TAKS</td>
<td>R</td>
<td>.70</td>
<td>.55</td>
</tr>
<tr>
<td>Scores</td>
<td>M</td>
<td>.56</td>
<td>.71</td>
</tr>
<tr>
<td>Grade 4</td>
<td>R</td>
<td>.65</td>
<td>.62</td>
</tr>
<tr>
<td>Class Grades</td>
<td>M</td>
<td>.52</td>
<td>.63</td>
</tr>
<tr>
<td>May</td>
<td>R</td>
<td>.76</td>
<td>.64</td>
</tr>
<tr>
<td>Renaissance</td>
<td>Benchmark M</td>
<td>.59</td>
<td>.71</td>
</tr>
</tbody>
</table>

**Concurrent Validity**

Concurrent validity coefficients from the data were product-moment correlation coefficients and the resulting associations were as follows:

- TAKS Reading with Teacher Grades in Reading in Grade 3, $r = 0.71$;
- TAKS Mathematics with Teacher Grades in Mathematics in Grade 3, $r = 0.69$;
- TAKS Reading with Teacher Grades in Reading at Grade 4, $r = 0.69$;
- TAKS Mathematics with Teacher Grades in Mathematics at Grade 4, $r = 0.58$;
- TAKS 3rd Grade Mathematics with ITBS Mathematics, $r = 0.66$;
- TAKS 3rd Grade Reading with ITBS Reading Comprehension, $r = 0.73$;
- TAKS Mathematics with Renaissance Benchmark scores at Grade 3, $r = 0.71$;
- TAKS Reading with Renaissance Benchmark scores at Grade 3, $r = 0.76$.

As shown, the lowest coefficient was 0.58 and the maximum was 0.76; therefore, all were within an acceptable range for concurrent validity standards.
Predictive Validity
Predictive validity coefficients for the TAKS were:
- TAKS Reading at Grade 3 with TAKS Reading at Grade 4, $r = 0.70$;
- TAKS Reading at Grade 3 with Teacher Grades in Reading at Grade 4, $r = 0.52$;
- TAKS Mathematics at Grade 3 with TAKS Mathematics at Grade 4, $r = 0.71$; and
- TAKS Mathematics at Grade 3 with Teacher Grades in Mathematics at Grade 4, $r = 0.63$.

As with concurrent validity coefficients, the predictive validity coefficients fell with an acceptable range and some were quite strong.

Construct Validity
Finally, construct validity was addressed based on theoretical underpinnings and the timing of assessments. The end-of-the-year Benchmark test, as are the TAKS, a standards-based test administered at the conclusion of the academic year. The 2nd Benchmark test is a standards-based test administered at mid-year. The ITBS is a content-based test administered at mid-year. The numerical grade given by the teacher is based on standards, content, and performance during the entire year. Finally, the 1st Benchmark test is a standards-based test administered at the beginning of the year. Table 3 shows the results for 3rd grade reading. As can be seen in Table 3, the predicted rank order of the correlations and the actual rank order of the correlations between TAKS 3rd grade reading and other assessments were analyzed using the Spearman’s rank order correlation ($\rho$). Only the second and third ranked values changed order from actual to hypothesized resulting in a Spearman’s correlation of $\rho = 0.90$. Consequently, this means that the agreement between the hypothesized pattern of correlations based on timing and foundational characteristics of the assessments was quite consistent with the actual empirical results. This provides some evidence of construct validity to the 3rd grade TAKS reading test.

Table 3
Construct Validity for TAKS 3rd Grade Reading Test

<table>
<thead>
<tr>
<th>Test/Assessment</th>
<th>Hypothesized Rank</th>
<th>$r$ with TAKS Reading</th>
<th>(rank) Rationale</th>
</tr>
</thead>
<tbody>
<tr>
<td>3rd Benchmark Test</td>
<td>1</td>
<td>0.76(1)</td>
<td>Both standards-based; both end-of-the-year</td>
</tr>
<tr>
<td>2nd Benchmark Test</td>
<td>2</td>
<td>0.71(3)</td>
<td>Both standards-based; 3 months difference</td>
</tr>
<tr>
<td>ITBS Reading Comprehension</td>
<td>3</td>
<td>0.73(2)</td>
<td>Standards vs Content Based; 3 months apart</td>
</tr>
<tr>
<td>Final Numerical Grade</td>
<td>4</td>
<td>0.71(4)</td>
<td>Standards vs Content Performance, Standards based; end-of-the-year vs on-going</td>
</tr>
<tr>
<td>1st Benchmark Test</td>
<td>5</td>
<td>0.68(5)</td>
<td>Both standards-based; Beginning of year vs end of the year</td>
</tr>
</tbody>
</table>

Spearman Rho Correlation = 0.90
Similarly, Table 4 displays the results for the third grade TAKS mathematics construct validity based on the same rationale used for reading. As shown, the validity coefficient defined by Spearman’s rho was 0.50, lower than for reading, but marginally acceptable.

**Table 4**

<table>
<thead>
<tr>
<th>Test/Assessment</th>
<th>Hypothesized Rank</th>
<th>( r ) with TAKS Reading</th>
<th>(rank) Rationale</th>
</tr>
</thead>
<tbody>
<tr>
<td>3rd Benchmark Test</td>
<td>1</td>
<td>0.71(2)</td>
<td>Both standards-based; both end-of-the-year</td>
</tr>
<tr>
<td>2nd Benchmark Test</td>
<td>2</td>
<td>0.74(1)</td>
<td>Both standards-based; 3 months difference</td>
</tr>
<tr>
<td>ITBS Reading Comprehension</td>
<td>3</td>
<td>0.66(5)</td>
<td>Standards vs Content Based; 3 months apart</td>
</tr>
<tr>
<td>Final Numerical Grade</td>
<td>4</td>
<td>0.69(4)</td>
<td>Standards vs Content Performance, Standards based; end-of-the-year vs on-going</td>
</tr>
<tr>
<td>1st Benchmark Test</td>
<td>5</td>
<td>0.69(3)</td>
<td>Both standards-based; Beginning of year vs end of the year</td>
</tr>
</tbody>
</table>

Spearman Rho Correlation = 0.50

**Conclusions and Recommendations**

The results of the present study provide some evidence of validity for the Texas high stakes test. The pattern of associations in the MTMM matrix was consistent with a high level of convergent and a quite acceptable level of discriminant validity. The various kinds of correlations did conform well with the pattern predicted by MTMM methodology. Therefore, the TAKS does a credible job of measuring mathematics and reading for 3rd and 4th grade students and the scores are not cluttered with extraneous domains irrelevant to mathematics and reading achievement.

Statistical validity in the form of concurrent and predictive validity results ranged from moderate to strong. Overall, the magnitude and direction of the correlations were within recommended ranges for validity coefficients. Construct validity based on the purpose and timing of the assessments provided good validity for reading, but only marginally adequate for mathematics. Overall, the TAKS stacked up with validity criteria quite well. If a common criteria for assessing validity coefficients for cognitive measures is used as a point of reference, such as \( r = 0.50 \), the TAKS results were generally well above the acceptable standard (McIntire & Miller, 2000, p. 157). From a psychometric perspective, and within the limitations of the present study, the TAKS seems to be measuring what it purports to measure. Even though strategically placed higher correlations would be preferred, the data show the TAKS to be sufficiently valid for assessing 3rd and 4th grade reading and mathematics achievement.
The results of the study coupled with the content validity information provided by the state show that the TAKS is sufficiently reliable and valid for assessing achievement in reading and mathematics at these grade levels. Other domains of achievement (i.e., social studies) and other grade levels were not addressed in the current study and generalizations to other academic fields and levels of study are not warranted. Because the study involved a single state, the degree of validity for application in other states cannot be established with the present information. Further, while the TAKS test seems to be psychometrically sound for school-wide assessment, it is not without weaknesses and should be used with other information when assessing individual knowledge, skills, and abilities.

As the present study examined validity using the TAKS scale score, nothing is implied regarding any cut-off scores for individuals or classifications of the achievement scores (i.e., Commended Performance, Met the Standard, Did Not Meet the Standard). Whether the value judgments imposed by categorizing the TAKS scores (e.g., Exemplary, Recognized, Academically Acceptable, etc.) have merit is a question beyond the scope of this study.

Despite the fact that test results may be misused in an infinite number of ways, careful and conservative interpretations of state-mandated TAKS mathematics and reading scores for third and fourth grade students can have some value. This value lies in predicting and confirming other educational achievement benchmarks thereby aiding in a diagnostic-prescriptive process of education.

High stakes decisions based on mandated tests are commonplace and their use will probably continue (Rose and Gallup, 2004). Consequently, psychometric studies of the validity and technical qualities of such tests in other subject areas, at various grade levels, and in other states should be rigorously pursued. Likewise, the results should be disseminated as advocated in the AERA/APA Standard for Educational and Psychological Testing. Whether such tests are to be challenged or supported, empirical evidence of the psychometric qualities constitutes a reasonable and minimal information base for the initiation of policy discussions.
References


