Construct Validity of the Spanish Version of a State-Mandated High-Stakes Test (TAKS)

Juanita M. Reyes & Dale M. Johnson
Tarleton State University

Abstract

State-level implementation of standardized testing for making important decisions has serious implications. Because of high-stakes consequences of the test results, the assessments need to satisfy technical psychometric standards. The Spanish version of the Texas Assessment of Knowledge and Skills (TAKS) is a state-mandated exam designed for Spanish-speaking students that has not been adequately evaluated for psychometric qualities. Three replicated studies were conducted using classical validation methodology to analyze psychometric properties of the TAKS exam for English language learners. Concurrent validity coefficients ranged from .52 to .78 and predictive validity coefficients ranged from .47 to .68 in reading and mathematics respectively. Theoretical expectations were verified by the multitrait-multimethod and factor analytic methodology. The study provided evidence of adequate construct validity for the Spanish version of the TAKS for assessing reading and mathematics achievement for elementary level English language learners. Although best practices in testing do not encourage high-stakes decisions based on a single measurement, the Spanish version of the TAKS demonstrated sufficient psychometric qualities for use as one indicator of achievement for elementary Spanish-speaking students.

The purpose of this study was to investigate the construct validity of the Spanish version of a state-mandated high-stakes test for a population of third and fourth grade Limited English Proficient (LEP) students. In response to the report A Nation at Risk (National Commission on Excellence in Education, 1983) in which concerns were expressed about the quality of public education, the educational reform effort known as No Child Left Behind Act (NCLB) became policy and defined accountability and assessment in public schools. Major components of NCLB are state-mandated written achievement tests accompanied by non-trivial consequences, and are therefore, frequently referred to as “high stakes” tests. While there is considerable policy debate about the role of high-stakes testing in schools, they are commonplace and such tests will continue to increase (Rose & Gallup, 2004).

The Standards for Educational and Psychological Testing (American Educational Research Association, 1999) states “... 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). Because individual test scores are frequently used as placement in academic remediation, for promotion from one grade to the next (McCombs, Kirby, Barney, Darilek, & Magee, 2005), and as a criterion for graduation (Kober, Chudowsky, Chudowsky, Gayler, & McMurrer, 2006), state-mandated standardized assessments should satisfy high-stakes criteria.

Not only does NCLB address school accountability measures, the act holds schools responsible for closing the long-standing achievement gap between minority and non-minority students (No Child Left Behind Act, 2001). Kindler (2002) reported results of a survey taken during the 2000-01 academic year showing 460 languages were spoken in the nation. The report documented that Spanish was the native language of the majority (79.2%) of English language learners (ELLs). Hispanics are the largest minority group consisting of over 42.7 million people in the United States and remains the fastest growing minority population (U. S. Census Bureau, 2006).
The rapid increase in the Hispanic population presents challenges to educators charged with satisfying educational needs of Spanish-speaking ELLs and this responsibility is complicated by the requirements of high-stakes testing. Abedi (2005) concluded that “Due to the rapid growth of the population of English language learners (ELLs) in the United States, and because of the confounding of language proficiency with content assessments of ELL students, issues concerning the assessment of these students deserve special attention” (p. 175). Abedi (2004) also noted that high-stakes testing of content knowledge for ELLs may result in depressed scores because the students cannot interpret the vocabulary and language used in the structure of the test. Research on assessment validity for ELL students has been limited (Baker & Good, 1995; McMaster, Wayman, & Cao, 2006) and doubt has been raised about whether tests function equally for ELL and non-ELL students (Bentz & Pavri, 2000). In order to reliably monitor the achievement gap between ELLs and native-speaking students, tests need to conform to high psychometric standards (Espin, Wallace, Campbell, Lembke, Long, & Ticha, 2008). While there is some evidence that state-mandated standardized tests have adequate psychometric qualities for English-speaking students (Burk, Johnson, & Whitley, 2005; Espin et al., 2008; Green, Winters, & Foster; 2003), the quality of the assessments for Spanish-speaking students is questionable (Kramer, Robertson, & Rodríguez, 2004; Guerrero, 2002) and lacking empirical psychometric investigation.

**Method**

*The Standards for Educational and Psychological Testing* (1999) clearly define the roles (pp. 9 – 24) and methods associated with convergent, discriminant, and test-criterion (concurrent and predictive) validity (p. 14) and content validity (p. 11). As achievement tests, standard testing practice described in virtually any elementary measurement text calls for presentation of evidence documenting various types of psychometric properties including validity (e.g., Linn & Gronlund, 2000). The present study does not address content validity or reliability of the Spanish version of the achievement test studied because the Texas Education Agency (2003) provides extensive and detailed descriptions of how the test is aligned with curriculum standards and subject matter content (content validity) as well as adequate coefficients of internal consistency reliability. Rather, the present study addresses various types of statistical validity including concurrent and predictive validity along with the more general and inclusive construct validity.

Campbell and Fiske (1959) provided procedures for logically studying patterns of correlation coefficients by analyzing a multitrait-multimethod matrix (MTMM). The use of the MTMM for studying convergent and discriminant validity has been widespread and used extensively for educational measures (Sawilowsky, 2002). Concurrent and predictive validity has traditionally been expressed as product-moment correlation coefficients (Linn & Gronlund, 2000) between test scores and a criterion score or performance.

In order to investigate the validity of the Spanish version of the state exam, three studies were conducted. Study 1 examined evidence for predictive validity and construct (discriminant) validity of the reading and mathematics tests. Study 2 examined the concurrent and construct (convergent and discriminant) validity of the exam. Finally, in keeping with the principle set forth by McIntire and Miller (2000) that “. . . investigations of validity include cross-validation” (p. 158), Study 3 served as a replication study to assess the degree to which the exam demonstrated concurrent, predictive, and construct validity. The state-mandated assessment which served as the focus of the study was the Spanish version of the *Texas Assessment of Knowledge and Skills* (TAKS).
**Study 1**

*Participants for Study 1*

Study 1 participants were Hispanic Limited English Proficient (N = 149) bilingual students in a southwest Texas city. Students were enrolled in 4th and 5th grade elementary classes and had attended the school since their first grade year. Data were provided by the school district’s central accountability office and consisted of 4th grade Spanish reading and mathematics TAKS scale scores and 5th grade English reading and mathematics scale scores for the same students. The entire sample was Hispanic in a bilingual education program and consisted of about half males (N = 74) and half females (N = 75).

*Design and Procedure for Study 1*

Predictive validity of the TAKS was investigated by correlating 4th grade Spanish reading and mathematics scale scores with the English version of TAKS one year later (5th grade). Thus, the 5th grade English version of reading and mathematics assessment served as the criterion for the predictive validity study. A product-moment correlation was calculated for reading and mathematics to serve as an indicator of predictive validity.

A modified MTMM matrix was used to examine patterns of correlation coefficients. Two traits were assessed (reading and mathematics) and one method (TAKS scale scores) was used. Therefore, the matrix consisted of Multitrait-Monomethod elements. The coefficients representing monotrait, monomethod elements in the matrix were compared with heterotrait, monomethod elements to assess discriminant validity. For Study 1, the degree to which the test shows discriminant validity depends on the extent to which monotrait, monomethod correlations are larger than heterotrait, monomethod correlations.

*Study 1 Results*

Predictive validity coefficients for the Spanish version of the TAKS were expressed as product-moment coefficients. Correlating Spanish TAKS reading scale scores from 4th grade with English version TAKS reading scale scores at 5th grade resulted in a correlation of \( r = .58 \). The 4th grade Spanish TAKS mathematics scale scores correlated with 5th grade mathematics scale scores resulting in \( r = .68 \).

The modified MTMM (multitrait, monomethod) matrix shows correlations between the same traits and between different traits. Monotrait, monomethod elements should show a stronger correlation than heterotrait, monomethod elements (Campbell & Fiske, 1959). As shown in Table 1, the monotrait correlations are .68 and .58, whereas the heterotrait coefficients were .37 and .58. The average size on the monotrait and heterotrait coefficients were determined by the method recommended by Bruning and Kintz (1997). Averaging two or more correlations requires transforming each \( r \) into a Fisher’s z-score before computing a mean. The z-scores are averaged and the average z-scores are converted back to product-moment \( r \)’s. The average size of the monotrait correlations was \( r = .63 \) whereas the heterotrait average \( r \) was \( r = .48 \). Thus, the results were consistent with theoretical expectations described by Campbell and Fisk (1959).
Participants for Study 2

Participants (N = 100) in Study 2 were 3rd grade students classified as Limited English Proficient (LEP) who had attended a southwestern urban elementary school since their first grade year. Data for male (N = 54) and female (N = 46) students in the sample consisted of 3rd grade Spanish version reading and mathematics scale scores and end-of-the-year teacher assigned numerical (percentage) grades based on performance in reading and mathematics classes.

Design and Procedure for Study 2

Data were provided by the school district’s accountability office and included Spanish version TAKS scale scores in reading and mathematics as well as percentage grades assigned by the 3rd grade teacher. Subscale scores representing objectives in the mathematics domain (6 objectives) and reading domain (4 objectives) were also collected.

Concurrent validity of the 3rd grade Spanish TAKS was assessed by correlating TAKS reading and mathematics scale scores with teacher assigned end-of-the-year grades. TAKS exams were administered at the conclusion of the school year and grades were assigned at the end of the year, but before the results of the TAKS were known. Therefore, because of the simultaneous timing of the two assessments the correlations represent concurrent validity coefficients.

Study 2 also utilized the MTMM matrix to investigate construct validity. The MTMM consisted of two methods (TAKS and grades) and two traits (reading and mathematics). Patterns of correlations representing monotrait, heteromethod elements and heterotrait, heteromethod elements were analyzed. According to logic provided by Campbell and Fiske (1959), strong monotrait, heteromethod coefficients support convergent validity whereas heterotrait, heteromethod coefficients should be the smallest coefficients in the matrix.

Additional evidence for construct validity was provided by a factor analysis of the reading and mathematics achievement subscale scores representing subject-matter objectives and teacher grades to determine the extent that the two subject matter domains (reading and mathematics) would be represented by a two-factor solution. A principal axis analysis extracted two factors and the factors were rotated using a varimax procedure to facilitate interpretation. A scree plot confirmed that a two-factor solution was reasonable based on a common criterion for limiting the number of factors as only two factors had eigenvalues greater than one.
Results of Study 2

Correlation coefficients between Spanish TAKS scale scores in reading and mathematics and teacher assigned grades in reading and mathematics were used to quantify concurrent validity. The correlation between TAKS reading scores at grade 3 and end-of-the-year teacher grades for grade 3 was $r = .64$. Similarly, TAKS mathematics scores correlated with teacher assigned mathematics grades $r = .73$.

The MTMM matrix in Table 2 shows the correlations between achievement test scores and teacher grades. The major diagonal coefficients are also concurrent and convergent validity coefficients and are represented by monotrait, heteromethod elements in the MTMM matrix. The average monotrait, heteromethod correlation (Bruning and Kintz, 1997) was $r = .69$ and the average heterotrait, heteromethod correlation was $r = .55$. The relative size of the average coefficients provides support for convergent and divergent validity (Campbell and Fiske, 1959).

<table>
<thead>
<tr>
<th>End-of-the-year Teacher Grades for Grade 3</th>
<th>3rd Grade Spanish-version TAKS</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mathematics</td>
<td>.73</td>
</tr>
<tr>
<td></td>
<td>Reading</td>
<td>.53</td>
</tr>
</tbody>
</table>

*Note: monotrait, heteromethod average $r = .69$
  heterotrait, heteromethod average $r = .55$*

Construct validity was investigated by a factor analysis to determine if common factors were consistent with subject matter as assessed by Spanish TAKS and teacher grades. Table 3 shows a two-factor structure representing reading and mathematics subject area domains.
The two factors accounted for about 60% of the total variance in the fourteen variables. The first factor was clearly a “mathematics” factor and accounted for just over 30% of the total variance. The second factor was a “reading” factor and accounted for just less than 30% of the total variance. Factor loadings (a.k.a. pattern/structure coefficients) greater than .55 were retained for interpretation and show strong relations between the specific assessments and the corresponding

<table>
<thead>
<tr>
<th>Variable</th>
<th>Factor I Mathematics</th>
<th>Factor II Reading</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mathematics Scale Score</td>
<td>.90</td>
<td>&lt;.55</td>
</tr>
<tr>
<td>TAKS Math Objective 1: Numbers, Operations, &amp; Quantitative Reasoning</td>
<td>.75</td>
<td>&lt;.55</td>
</tr>
<tr>
<td>TAKS Math Objective 4: Concepts and Uses of Measurement</td>
<td>.74</td>
<td>&lt;.55</td>
</tr>
<tr>
<td>TAKS Math Objective 6: Mathematical Processes and Tools</td>
<td>.74</td>
<td>&lt;.55</td>
</tr>
<tr>
<td>Teacher Assigned End-of-Year Mathematics Grades</td>
<td>.66</td>
<td>&lt;.55</td>
</tr>
<tr>
<td>TAKS Math Objective 2: Patterns, Relationships, &amp; Algebraic Reasoning</td>
<td>.65</td>
<td>&lt;.55</td>
</tr>
<tr>
<td>TAKS Math Objective 3: Geometry and Spatial Reasoning</td>
<td>.57</td>
<td>&lt;.55</td>
</tr>
<tr>
<td>TAKS Math Objective 5 Probability and Statistics</td>
<td>.55</td>
<td>&lt;.55</td>
</tr>
<tr>
<td>Reading Scale Score</td>
<td>&lt;.55</td>
<td>.91</td>
</tr>
<tr>
<td>TAKS Reading Objective 1: Basic Reading Understanding</td>
<td>&lt;.55</td>
<td>.81</td>
</tr>
<tr>
<td>TAKS Reading Objective 2: Applying Knowledge of Literary Elements</td>
<td>&lt;.55</td>
<td>.76</td>
</tr>
<tr>
<td>TAKS Reading Objective 3: Using Strategies to Analyze</td>
<td>&lt;.55</td>
<td>.73</td>
</tr>
<tr>
<td>Teacher Assigned End-of-Year Reading Grades</td>
<td>&lt;.55</td>
<td>.70</td>
</tr>
<tr>
<td>TAKS Reading Objective 4: Applying Critical-Thinking Skills</td>
<td>&lt;.55</td>
<td>.65</td>
</tr>
</tbody>
</table>

_Note:_ Factor Loadings of 0.55 or greater were retained for interpretation.
latent variables representing the subject matter areas. The factors show that the Spanish math TAKS assessments and classroom performance converge into the homogeneous mathematics dimension while reading scores and performance converge into the reading dimension.

**Study 3**

*Participants for Study 3*

Participants in the replication study (Study 3) consisted of Limited English Proficient (N = 166) bilingual students in the 4th and 5th grade elementary school in a southwest urban district. All participants had been enrolled in the same school since the first grade and consisted of more male students (N = 91) than female (N = 75).

*Design and Procedure for Study 3*

An IRB form was approved by the participating school district and the accountability office of the district provided data for the analyses. Data collected on the same set of students consisted of 4th grade Spanish-version TAKS reading and mathematics scale scores; 5th grade English-version of TAKS reading and mathematics scale scores; 4th grade end-of-the-year teacher numerical (percentage) scores for mathematics and reading; and 5th grade end-of-the-year teacher numerical grades for math and reading.

Predictive validity of the Spanish-version was reexamined by correlating scale scores on the Spanish version of the TAKS with performance on the English version one year later – the English version serving as the criterion. A MTMM was examined to determine convergent and discriminant validity exhibited by the data. The two traits in the matrix were reading and mathematics and the two methods of assessment were TAKS scale scores and teacher end-of-the-year numerical grades. Concurrent validity of the Spanish version of the TAKS exam was addressed by correlating TAKS reading and mathematics scale scores from the Spanish version with the teacher numerical grades in mathematics and reading.

*Results of Study 3*

Predictive validity was quantified by a Pearson product-moment correlation coefficient between 4th grade Spanish TAKS reading and mathematics scale scores and the criteria in the form of 5th grade English version TAKS scale scores in the same subject area domains. Also, predictive validity was examined by correlation between Spanish TAKS scores from the 4th grade with teacher numerical grades given at the end of the 5th grade. Predictive validity coefficients were as follows:

- 4th grade Spanish-version TAKS reading with 5th grade English-version TAKS reading, $r = .50$;
- 4th grade Spanish-version TAKS mathematics with 5th grade English-version TAKS mathematics, $r = .53$;
- 4th grade Spanish-version TAKS reading with 5th grade Teacher grades, $r = .47$; and
- 4th grade Spanish-version TAKS mathematics with 5th grade Teacher grades in mathematics, $r = .53$. 
Concurrent validity was assessed by correlating 4th grade Spanish TAKS scale scores with teacher numerical grades assigned at the end of the 4th grade. Concurrent validity coefficients were as follows:

- 4th grade Spanish-version TAKS reading scale scores with end-of-the-year grades in reading, $r = .54$;
- 4th grade Spanish-version TAKS mathematics scale scores with end-of-the-year grades in mathematics, $r = .60$.

Elements of the MTMM were Pearson correlation coefficients among 4th grade Spanish-version TAKS scale scores in reading and mathematics, 5th grade English-version TAKS scale scores in reading and mathematics, 4th grade end-of-the-year teacher grades in reading and mathematics, and 5th grade end-of-the-year grades in mathematics and reading coursework. Table 4 shows the MTMM elements.

**Table 4**

<table>
<thead>
<tr>
<th></th>
<th>4th Grade</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Spanish-Version TAKS</td>
<td>End-of-the-Year Grades</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5th Grade TAKS</td>
<td>Reading</td>
<td>Mathematics</td>
<td>Reading</td>
<td>Mathematics</td>
<td></td>
</tr>
<tr>
<td>English-version</td>
<td>.50(^a)</td>
<td>.46(^c)</td>
<td>.52(^b)</td>
<td>.40(^d)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.45(^c)</td>
<td>.53(^a)</td>
<td>.57(^d)</td>
<td>.63(^b)</td>
<td></td>
</tr>
<tr>
<td>5th Grade Teacher Grades</td>
<td>Reading</td>
<td>Mathematics</td>
<td>Reading</td>
<td>Mathematics</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.54(^b)</td>
<td>.54(^d)</td>
<td>.72(^a)</td>
<td>.56(^c)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.49(^d)</td>
<td>.60(^b)</td>
<td>.67(^c)</td>
<td>.63(^a)</td>
<td></td>
</tr>
</tbody>
</table>

**Note:**

- \(^a\) monotrait, monomethod average $r = .60$
- \(^b\) monotrait, heteromethod average $r = .57$
- \(^c\) heterotrait, monomethod average $r = .54$
- \(^d\) heterotrait, heteromethod average $r = .50$
- \(^a\) and \(^b\) are convergent predictive validity coefficients and are larger.
- \(^c\) and \(^d\) are discriminant validity indicators and are smaller.

The MTMM consisted of two traits (reading and mathematics) and two methods of assessment (TAKS examination and teacher administered end-of-the-year grades). The coefficients superscripted with “a” are monotrait, monomethod coefficients reflecting the fact that the same trait is measured by the same method. The monotrait, monomethod coefficients should be the largest coefficients in the matrix. The coefficients with “b” superscripts are monotrait, heteromethod elements representing correlations between the same traits measured with different methods. They are hypothesized to be the second largest coefficients in the matrix and are indicators of convergent validity. Elements with “c” superscripts are heterotrait, monomethod because they are different traits measured with the same method and are hypothesized to be smaller than the coefficients labeled “a” and “b”. Finally, the smallest values in the MTMM are expected to be the coefficients superscripted as “d”; heterotrait, heteromethod. As shown in the MTMM, evidence of construct validity is confirmed by the fact that the average of the monotrait elements ($r = .59$) is
higher than the average of the heterotrait elements ($r = .52$). Consistent with Campbell and Fiske's (1959) expectations the average correlations were:

average “a” element (monotrait, monomethod) was the highest average, $r = .60$;
average “b” element (monotrait, heteromethod) which are convergent validity coefficients was second largest, $r = .57$;
average “c” element (heterotrait, monomethod) was third largest, $r = .54$; and
average “d” element (heterotrait, heteromethod) was the smallest, $r = .50$.

**Summary and Discussion**

The purpose of this study was to evaluate statistical validity of the Spanish-version of a state-mandated high-stakes test for LEP students. Findings provide some evidence in support of the Spanish-version of the TAKS reading and mathematics exam for elementary school students. A summary of the results is shown in Table 5. Given a criterion that “validity coefficients rarely exceed .50” (McIntire & Miller, 2000, p. 157), statistical validity demonstrated by concurrent and predictive validity coefficients range from moderate to strong. Patterns of correlations in the MTMM matrices were consistent with theoretical expectations and conformed well to the pattern predicted by the MTMM methodology. Factor analysis confirmed that the measures converged into distinct latent traits representing mathematics and reading achievement.

To the extent that the state-mandated test is aligned with the classroom instruction (curricular validity), educators can place fairly strong confidence in the Spanish-version of the TAKS scores as measures of achievement in elementary school reading and mathematics. The results also supplement and expand validity information provided by Burk, Johnson, and Whitley (2005) and Mahon (2006). The study also moderates concerns regarding the effects of the transadaptation process expressed by Kramer, et al. (2004) and Guerrero (2002). However, while the evidence supports moderate to high validity of the Spanish-version of the test, there is room for improvement. As recommended by American Educational Research Association (AERA,2000), high-stakes decisions should not be made exclusively on the basis of the TAKS or any other single measure. With respect to the Spanish-version of the TAKS, teachers, administrators, parents, students, and other educational stakeholders can have confidence that the elementary level reading and mathematics test for 3rd and 4th grade students can play an important role as one source of information in assessing overall achievement.
References


