

Chapter 17: Validity

Overview

In proper usage, one is interested in making proper interpretations of a test score, so test makers are responsible for accumulating evidence that support the intended interpretations and uses of the scores (Kane, 2006). In the case of Texas Assessment of Knowledge and Skills (TAKS), TAKS–Modified (TAKS–M), and end-of-course (EOC) assessment, results are used to make inferences about students’ knowledge and understanding of the Texas Essential Knowledge and Skills (TEKS). Texas English Language Proficiency Assessment System (TELPAS) assessment results provide a measure of progress in English language acquisition in alignment with the English language proficiency standards (ELPS) in the TEKS. This chapter provides evidence about the processes used to develop the tests and the analytic studies conducted to better understand interpretations that may be assigned to individual test scores. The search for validity evidence is a never-ending process, and future technical reports will include additional information in this regard.

Evidence Based on Test Content

Standards-referenced assessments, such as the TAKS, TAKS–M, TELPAS, and EOC are based on an extensive definition of the content they assess. Test validity is, therefore, content based and tied directly to the statewide curriculum. To ensure the highest level of content validity, the process of aligning TAKS, TAKS–M, TELPAS, and EOC to the curriculum was carefully implemented and included review by numerous committees of Texas educators. The federal No Child Left Behind Act (NCLB) requires the states to verify the alignment of their assessment tests to the adopted curriculum.

When TAKS was designed as the standards-referenced assessment for the TEKS, advisory committees consisting of educators from districts across the state were formed for each subject area at each grade level. Teachers, test development specialists, and Texas Education Agency (TEA) staff members worked together in these committees to identify TEKS student expectations that were important to assess and to develop test objectives, item development guidelines, and test-item types. In addition, starting in 2001–2002, committees convened to review and edit TAKS items for content and bias and to review data from field testing. A similar process was conducted for TAKS–M, TELPAS, and EOC when they were developed. For TAKS–M, advisory committees meet to review the original TAKS item and the modified version of the item as it would appear on TAKS–M. These committees confirmed that the modified version of the item still measures the TEKS student expectation it was measuring as a TAKS item. The committees also review and edit the modified items for content and bias.

Relationship to the Statewide Curriculum

The item writers and reviewers for each stage of development verify the alignment of test items with the objectives to ensure that the items measure appropriate content. The sequential stages of item development and item review provide many opportunities for Texas educators to offer suggestions for improving or eliminating items and to offer insights into the interpretation of the statewide curriculum. The nature and specificity of these various review procedures provide additional strong evidence for the content validity of the TAKS, TAKS–M, TELPAS, and EOC assessments.

Educator Input

Texas educators provide valued input on the content and the connection between the items and the statewide curriculum. Many current and former Texas educators, and some educators from other states, work as independent contractors to write items specifically to measure the objectives. These multiple sources of expertise provide for a system of checks and balances for item development and review that reduces single-source bias that might be introduced if items were written by a single author. In other words, because test items are written by many different people with diverse backgrounds, it is less likely that items will suffer from a bias that might occur if items were written by a single author. These multiple sources of expertise provide for a system of direct input from educators which offers additional evidence regarding the validity of constructed TAKS, TAKS–M, TELPAS, and EOC tests.

Test Developer Input

The staff at TEA, as well as professional test developers from Educational Testing Service (ETS), Pearson, and Questar, Inc., provide a wealth of test-building experience, including content expertise. Each internal review of an item by these experts increases the probability of the item being an accurate measure of the intended objective. Hence, these reviews are offered as additional evidence for the content validity of the TAKS, TAKS–M, TELPAS, and EOC assessments.

Test Expert Input

TEA, in conjunction with Pearson, receives ongoing input from a panel of national testing experts regarding all plans for collecting validity evidence for the Texas assessments. In the case of the state's English language learner (ELL) student assessment procedures, language acquisition and psychometric experts are involved in all phases of assessment development and refinement. Several times a year, Texas convenes an ELL assessment focus group to provide input as a team of language acquisition experts. This group provides input on technical and practical issues related to the development and refinement of TELPAS and linguistically accommodated testing (LAT) procedures required under NCLB. The ELL assessment focus group is composed predominantly of bilingual and English as a second language (ESL) educators and coordinators, district testing coordinators, and campus administrators. Texas has also elicited psychometric input from

its Technical Advisory Committee (TAC) on ELL assessment issues. The Texas Technical Advisory Committee (TTAC) has provided input on several plans for gathering reliability and validity evidence for TELPAS and has helped shape the plans for Spanish TAKS standard setting and TELPAS audits, both small and large-scale audits.

Evidence Based on Relationships to Other Variables

Another way to provide validity evidence is by analyzing the relationship between test performance and performance on some other measure. This other measure can be evaluated concurrently or in the future and is then correlated with the test score. In this way, the test score is compared with a criterion that is thought to be a reasonable estimate of the same construct the original test purports to measure. As part of the TAKS Higher Education Readiness Component, a concurrent validity study was conducted in 2004–2005 to correlate performance on exit level TAKS with performance on national testing programs.

TAKS

Criterion-related evidence of validity for TAKS was provided in a study conducted by TEA and Pearson to fulfill the Senate Bill 103 requirement that TEA implement a college readiness component as part of the TAKS. The research, called the Higher Education Readiness Component study, included two parts: a contrasting groups study and a performance data correlation study. The contrasting groups study examined the performance of high school juniors on the first administration of the TAKS exit level mathematics and English language arts tests in 2003 as compared to performance on the same TAKS assessments by a sample of second semester college freshmen who had demonstrated college readiness through successful completion of their first semester courses.

The performance data correlation study examined student performance on TAKS in relation to performance on three college readiness measures used statewide for making college readiness and placement decisions: the Texas Academic Skills Program (TASP), the American College Test (ACT), and the Scholastic Assessment Test I (SAT I). The TAKS to TASP, TAKS to ACT, and TAKS to SAT I comparisons incorporated data collected from Texas public high school juniors who took the exit level TAKS and one or more of these other assessments in spring 2003. ACT and SAT I data also were collected for high school juniors who took the TAKS in spring 2004.

Results of the study indicated that the TAKS scale scores at the Met Standard performance level predicted ACT scale scores of approximately 20 for mathematics. Based on a national study of high school graduates from 2002 to 2004, 50% of students scored at or above this ACT score. The TAKS scale scores at the Met Standard performance level predicted ACT scale scores of approximately 18 for English. Of the high school students in the ACT data, 67% scored at least this high on the ACT English test.

Results of the study also indicated that the TAKS scale scores at the Commended Performance level predicted ACT scale scores of approximately 27 for mathematics. Based on a national study of high school graduates from 2002 to 2004, 15% of students scored at or above this ACT score. The TAKS scale scores at the Commended Performance level predicted ACT scale scores of approximately 24 for English. Of the high school students in the national study, 29% scored at least this high on the ACT English test.

Results of the study indicated that the TAKS scale scores at the Met Standard performance level predicted an SAT I scale score of approximately 470 for mathematics. Based on a national study of high school graduates, 50% of students scored at or above this SAT I score. The TAKS scale scores at the Met Standard performance level predicted an SAT I scale score of approximately 460 for English. Based on a national study of high school graduates, 50% of students scored at or above this SAT I score.

Results of the study indicated that the TAKS scale scores at the Commended Performance level predicted an SAT I scale score of approximately 620 for mathematics. Based on a national study of high school graduates, 25% of students scored at or above this SAT I score. The TAKS scale scores at the Commended Performance level predicted an SAT I scale score of approximately 580 for English. Based on a national study of high school graduates, 25% of students scored at or above this SAT I score. For further information about the study, see the “Higher Education Readiness Study” report in the 2007 Texas Education Agency Technical Report Series which can be found at <http://www.tea.state.tx.us/student.assessment/resources/techdigest/>.

Another source of criterion-related validity evidence for the TAKS tests is the Grade Correlation Study. This study compared the pass/fail rates of Texas students on the TAKS tests with their passing credit/not passing credit rates in their past related courses. Results indicated that a high percentage of students who pass the TAKS tests also pass their related courses. Small percentages of students passed the TAKS tests but did not pass their related courses, passed their related courses but did not pass the TAKS tests, or failed to pass the TAKS test or their related courses. For more details on the study, see the “Grade Correlation Study” report in the 2008 Texas Education Agency Technical Report Series which can be found at <http://www.tea.state.tx.us/student.assessment/resources/techdigest/>.

The transadaption process is an important method used to develop Spanish TAKS reading, mathematics, and science. In this process, items for Spanish TAKS were translated from English TAKS and adapted as necessary to ensure linguistic and cultural appropriateness.

A study conducted by Pearson (Davis, O’Malley & Wu, 2007) on the measurement equivalence of transadapted reading and mathematics tests provided an evidence of validity of TAKS tests. The study investigated the fit of a confirmatory factor model to student data from 2006 for grades 3 and 5 TAKS reading and grades 4 and 6 TAKS mathematics tests across English anchor forms and Spanish transadaptions.

The results of the study indicated good model fit within and across English and Spanish forms for both subjects and grades. The study suggested that the Spanish TAKS and English TAKS items function similarly and that the combination of transadapted and independently developed items on the Spanish TAKS tests is appropriate for assessing both reading and mathematics. The results of the study support that both English- and Spanish-versions of TAKS tests measure the same construct.

Gathering Validity Evidence for TAKS–Alt

As with TAKS, TAKS–Alternate (TAKS–Alt) test results are used to make inferences about students’ knowledge and understanding of the TEKS. Unlike other statewide assessments in Texas, TAKS–Alt is not a traditional paper-and-pencil or multiple-choice test. Instead, the assessment involves teachers observing students as they complete instructional activities that link to the grade-level TEKS curriculum. As mentioned in the introductory paragraph of this chapter, validity for this assessment is a process of collecting evidence to support inferences made from the scoring results, but different approaches than those used for TAKS have been used for TAKS–Alt.

Evidence Based on Test Content

Content validity evidence has been collected at all stages of the test development process. Evidence based on test content is information that shows the relationship between content and the construct measured by the test. TAKS–Alt was developed to align with the content defined by TEKS. An explicit mapping of the alignment of the TAKS–Alt with the TEKS can be seen in the following documents developed for TAKS–Alt: the TEKS Vertical Alignment documents, the TEKS Curriculum Framework for the Alternate Assessment documents, and the Example Instructional Activities documents. These documents can be found online at <http://www.tea.state.tx.us/student.assessment/resources/techdigest/>.

Content experts and special education experts have been involved in the development and refinement of the assessment since its inception. Focus groups consisting of teachers and other experts have provided reviews and feedback at multiple points throughout the development process. Educator advisory meetings were convened to provide feedback on the TAKS–Alt participation guidelines, the prototype assessment, and the scoring rubric. Educator review meetings provided teacher feedback on the alignment of the general education curriculum and access activities to the grade level TEKS curriculum and TAKS objectives. Volunteer teachers used the TAKS–Alt scoring rubric and provided feedback. The steering committee convened three times to provide feedback on all aspects of the assessment.

In addition, all internal meetings, steering committee meetings, educator advisory meetings, and educator review meetings are documented. All versions of TAKS–Alt related materials are being maintained, and a technical report is being written to describe the progression of the TAKS–Alt development. The technical report will include information collected during the activities described below.

TAKS–Alt Audit

Additional content validity evidence for the TAKS–Alt was obtained through a series of audits of student responses. Auditors reviewed 10% of the student folders consisting of responses to the four essence statements assessed during the spring 2007 field-test administration. This audit provided validity evidence for both the field-test administration and the 2007–2008 operational administration of the TAKS–Alt. As part of this review, the teacher-developed instructional activities were analyzed to evaluate how well the activities matched the assessed objective and how well the activities matched the assessed essence statement. Since the essence statements were developed to be the core of the knowledge and skill statement under each objective, it was expected that if there was a strong link between the activity and the essence statement, that link would also be present between the activity and the objective. Auditors also reviewed the primary documentation provided by teachers to support the scoring of the instructional activity for each of the four essence statements. Auditors were shown the teacher ratings and evaluated how well the documentation matched the teacher ratings (or student score) for the instructional activity.

Auditors were asked to indicate their level of agreement (strongly agree, agree, disagree, or strongly disagree) with the following three statements for each student folder they viewed:

- The instructional activity is linked to the essence statement being assessed.
- The instructional activity is linked to the objective statement being assessed.
- The documentation supports the student’s score.

A total of 1879 student folders were viewed by 124 auditors. In terms of the teacher-developed instructional activity being linked to the essence statement, 95% of auditors agreed or strongly agreed with the statement. As expected, 95% of auditors also agreed or strongly agreed with the statement that the instructional activity linked to the objective being assessed. These results provide good validity evidence that the activities teachers are developing to assess students are linked to the content that the teacher planned to assess.

Responses regarding how well the provided documentation supported the student’s score were slightly less strong with 75% of auditors agreeing or strongly agreeing with the statement. Further training on student documentation and its link to the student’s score may result in more auditors agreeing that the documentation supports the student’s score. Training with this focus was available for the first operational administration of TAKS–Alt in 2007–2008.

The next TAKS–Alt audit is scheduled for the 2008–2009 school year. Adjustments may be made to the upcoming audit to collect additional validity evidence from that described above.

Evidence Based on Consequences of Testing

Validity evidence that shows the TAKS–Alt is having a positive impact on student learning and instruction has been collected through teacher surveys after the 2007–2008 test administration. Teachers generally agreed that students will be adequately prepared for the 2007–2008 administration. The majority of surveyed teachers were somewhat or very comfortable conducting instructional activities. Additionally, teachers were confident about the rating their students earned. Those teachers that participated in the pilot reported administering the pilot test to their students increased their own understanding of how to administer the field test.

For further information on the TAKS–Alt survey results see the 2007–2008 “TAKS–Alt Technical Report” in the 2008 Texas Education Agency Technical Report Series which can be found at <http://www.tea.state.tx.us/student.assessment/resources/techdigest/>.

Gathering Validity Evidence for TELPAS

The results of TELPAS assessments are used to make inferences about the progress that English language learners (ELLs) make in acquiring the English language. TELPAS results are used for both federal and state accountability purposes, as well as for monitoring performance and informing parents of their children’s English language proficiency level. Evidence supporting the validity of the reading, writing, listening, and speaking domains of TELPAS has been collected since the first administration in 2003–2004 and continues to be collected. Validity evidence has been gathered in different manners since multiple-choice tests are used to assess reading in grades 2–12, whereas holistically rated assessments are used to measure reading in K–1 and listening, speaking, and writing in K–12. In addition to the studies mentioned here, a wide range of validity studies and analyses have been conducted and documented in the Technical Report Series and Technical Digest for previous years. These documents are available on the TEA website at <http://www.tea.state.tx.us/student.assessment/resources/techdigest/>

Evidence Based on Test Content

TELPAS measures student performance in direct alignment with the English language acquisition skills and proficiency level descriptors defined by the Texas English language proficiency standards (ELPS), which are part of the Texas Essential Knowledge and Skills (TEKS) curriculum. The ELPS outline the instruction that ELLs must receive to support their ability to develop academic English language proficiency. Revised ELPS were approved by the State Board of Education in December of 2007. TELPAS assesses the ELPS for listening, speaking, reading, and writing.

Test Design

The multiple-choice TELPAS reading tests for grades 2–12 are designed to assess English language reading proficiency in a manner that provides meaningful diagnostic

information about how well ELLs are learning to read the English they need for academic success in U.S. schools. The test is built using four levels, or degrees, of linguistic accommodation, addressing the gradually reduced degree of linguistic accommodation that ELLs need as they progress from knowing little or no English to becoming fluent English readers. Four English language proficiency levels are reported: Beginning, Intermediate, Advanced, and Advanced High. Each proficiency level defined in the ELPS is characterized by the degree of linguistic accommodation that students at that level need to read English with understanding.

Each reading selection and test question is written to reflect a particular proficiency level associated with a particular degree of linguistic accommodation. The test blueprints require a certain number of test questions per proficiency level and per test objective (skill category). The score reports inform teachers of how successfully students demonstrate basic comprehension and analytical reading skills at the four proficiency levels. The content validity of the TELPAS reading assessment is supported by its test design in that it provides staged linguistic accommodations commensurate with second language learning as it measures reading skills that students need for academic success in all subject areas. The staged linguistic accommodation test design is shown below.

Figure 10. Staged Linguistic Accommodation Test Design

↑	TELPAS Reading Levels	Degree of Linguistic Accommodation / Key Features	
	Advanced High	Minimal	Minimal linguistic accommodation; texts highly comparable to those written for native English speakers
	Advanced	Moderate	Occasional picture support; contextual aids and organizational features support comprehension of longer texts on both familiar and unfamiliar language arts and content area topics
	Intermediate	Substantial	Frequent picture support; short texts written primarily on familiar topics; commonly used, everyday English and routine academic English
	Beginning	Extensive	Maximum picture support; basic comprehension of words, sentences, and short paragraphs; high-frequency, concrete English

Like the reading tests, the TELPAS holistically rated components are also aligned with the ELPS skills and proficiency level descriptors, and are designed to assess the English communication skills that ELLs need to engage meaningfully and successfully in learning the academic knowledge and skills required by the state. The holistically rated assessments draw upon second language acquisition research, research-based standards, the experience of Texas practitioners, and observational assessment practices.

The TELPAS holistically rated components are based on ongoing observations of the ability of ELLs to understand and use English during the very grade-level core content area instruction that is required by the state-mandated curriculum and assessed on the

state-mandated assessments. The proficiency level descriptors (PLDs) used for the assessment are the same as those in the ELPS. The PLDs form research-based English language proficiency continuum widely accepted by Texas practitioners who work daily with ELLs. As is typical of holistically scored assessments, students are evaluated on their overall performance in a global and direct way. The goal of NCLB English language proficiency assessments is to effectively assess the extent to which ELLs are making progress in and attaining academic language proficiency so that they can achieve their full academic potential. The TELPAS holistically rated assessments are designed to meet this goal head on. The fact that the results are based on a direct measure of the student's actual abilities rather than on an indirect measure provides strong content validity evidence.

Test Development and Construction

The process used to develop the multiple-choice TELPAS reading tests and TAKS tests is the same. The process adheres to The Standards for Educational and Psychological Testing (AERA/APA/NCME, 1999), is grounded in the state's ELP standards, and is guided by assessment experts and educators who have first-hand knowledge of the standards and students. As with TAKS, the TELPAS reading test construction process involves multiple reviews by both content and psychometric experts. The fact that the state follows the same thorough development process for the TAKS and TELPAS reading tests—and includes TAKS assessment and content area experts throughout the TELPAS reading development process—supports the content validity of TELPAS and its link to the state's academic content standards.

The state's decision to implement holistically rated TELPAS assessments stemmed from the need to respond to concerns related to excessive testing and field testing given the comprehensive assessment systems that were already in place and to avoid logistically impractical speaking and listening assessments given the state's high ELL population. The TELPAS holistically rated assessments address these concerns and uphold the commitment of the state to administer valid and reliable assessment instruments that support sound instructional practices and are appropriate for use in accountability systems. Because of the direct involvement teachers across the state have in the assessment process, the holistically rated assessments have had a direct and significant positive effect on classroom instruction.

More details about the development of TELPAS reading and holistically rated assessments can be found in the TELPAS chapter (Chapter 4, TELPAS).

Training and Administration Procedures

Evidence of the validity of TELPAS is supported by comprehensive training and administration procedures which ensure that teachers are prepared to perform their duties and district administrators follow procedures to ensure the integrity of the test administration.

Stringent training requirements are used to maximize rating accuracy associated with the holistically assessed components of TELPAS. These requirements include annual TELPAS rater training for all language domains. Raters must also complete a qualifying component of their training. Raters initially receive both online training and face-to-face instruction delivered by trainers who are trained directly by TEA. Training participants use the ELPS proficiency level descriptors (PLDs) as rubrics to practice rating Texas students shown in video segments and to practice rating authentic student writing collections. As the culminating component of their training, participants rate student performance online, and those who rate students with the required level of accuracy receive a rater qualification certificate. For previously qualified TELPAS raters, the state uses online refresher training courses that ensure continued rating accuracy. The annually required courses not only refresh and recalibrate raters in all language domains but serve as a vehicle to gradually enhance Texas teachers' knowledge base of second language acquisition processes. In spring 2008 alone, approximately 110,000 online training courses were completed and approximately 28,000 qualification certificates were awarded. The percent of teachers meeting rater qualification requirements has been 90% or higher each year.

In building the rater training systems, survey input was collected from training participants as part of the validation process. Feedback on the assessments, PLDs, and training itself was used in the development and enhancement of the training system. The TELPAS training system meets the state's goal of having both a valid and authentic assessment and a critical ongoing professional development tool that supports effective instruction so that teachers better understand and meet the educational needs of ELLs.

To promote consistency in the administration of TELPAS, the state has also implemented specific administration and validity and reliability procedures delineated in administration manuals. All testing personnel and principals are required to receive annual training on the contents of the manuals. They sign oaths certifying that they have been trained and understand their obligations in administering the tests and maintaining test integrity, security, and confidentiality. In addition, campus administrators are required to document the procedures they follow to support validity and reliability during the assessments, rater information is collected on each student's answer document, and principals sign the rating sheet of each rater in affirmation of oversight of the assessment process.

TELPAS Audits

Additional evidence supporting content and construct validity of TELPAS comes from annual audits. Since the 2004–2005 school year, the Texas Education Agency has conducted annual audits of the TELPAS assessment processes. Sampling procedures have been used to require district and campus testing personnel and raters to respond to audit questionnaires about assessment procedures followed and the quality of the training received for each language domain. Information about TELPAS audits is also included in Chapter 16: Reliability because the same crucial pieces of evidence support both the validity and reliability of TELPAS holistically rated components.

Writing Audits

The TELPAS writing assessments require trained raters to use the writing PLDs from the ELPS and student writing from classroom assignments to assign student English language proficiency levels. For four years, TEA has required sampled school districts to submit to the state the student writing collections they rated for the writing component of the assessment. Audit raters trained by the Pearson Performance Scoring Center re-rated the writing collections to monitor how well district raters were applying the PLDs as holistic scoring rubrics during the live assessment. The writing collections themselves were also examined to ensure that they contained the necessary types and number of student writing assignments, as stipulated in the administration manuals.

The first audit, conducted in 2005, was relatively small, and had the objective of helping the state establish and improve audit procedures. The second audit, conducted in 2006, was a larger study in which information from a large, representative number of districts and students was collected. This audit provided results for regional education service centers (ESCs) and large districts referred to as training entities because of their role in directly providing TELPAS training to teacher raters. The 2006 audit results provided reliability and validity evidence supporting the accuracy of teacher ratings for the writing domain. The spring 2007 TELPAS audit was smaller and served to provide ongoing evidence of validity and reliability at the level of the state rather than at the training entity level.

The results of the last three years of audits are presented in Table 14. The results of the 2008 writing audit provided evidence of rater accuracy at a level similar to that reported in the 2007 audit. The overall perfect agreement rate of 79% was found to be satisfactory based on the Pearson Performance Scoring Center ISO Standards, and the adjacent agreement rate was 98%. In addition, the presence of a high correlation ($r = 0.89$) and a high weighted kappa ($w = 0.81$) value underscored the strong agreement between the state auditor ratings and the teacher ratings.

Table 14. Writing Audit Results from 2006 to 2008

Year	Sample Size	Perfect Agreement Rate	Correlation Between District and State Raters	Kappa
2006	13357	77%	0.87	0.79
2007	542	76%	0.87	0.78
2008	1932	79%	0.89	0.81

Table 14 illustrates consistently high rating accuracy across multiple years and provides evidence of stability in interrater agreement over time. Since TELPAS scores across years are used in reporting student progress in language acquisition, the state's finding of rater accuracy over time supports inferences about annual student progress from TELPAS scores.

The state will continue to audit the writing process periodically to provide ongoing monitoring of rating effectiveness and to give district personnel feedback to support the administration of this assessment.

Listening and Speaking Audits

The first small-scale pilot audit of TELPAS listening and speaking components was conducted in spring 2008. In this small-scale study, audit raters with previous TELPAS rater training were provided a face-to-face refresher training in May 2008 and then sent to school districts to re-rate ELL students who had been assessed in March–April 2008 in their natural classroom settings.

The pilot encompassed two TELPAS grade clusters, grade 2 and grades 6–8, and included 43 internal (from TEA and Pearson) and external (from districts and education service centers) audit raters. Each audit rater was given a list of several students to rate blindly over the course of three days. The objectives of this small pilot were to examine the viability of this method of auditing rating efficacy and the feasibility of conducting this type of audit on a larger scale.

The audit raters received face-to-face training in two main areas: rating skills and audit logistics. The audit raters received refresher training in the proficiency level rubrics in order to observe students, assign ratings, document reasons for their ratings, and determine whether the students fell into the early, middle, or late stage within the proficiency levels assigned.

The pilot enabled audit raters to re-rate approximately 150 students in each of the two grade clusters. Although the pilot was small in scope and used only volunteer audit raters, students were sampled in a way that maximized representation of different regions of the state to the extent possible.

Questionnaire information from the audit raters was collected along with their ratings of students. The questionnaire results indicated that the majority of audit raters (60%) were able to rate 1–2 students per day if the students belonged to different classrooms at the same campus and 3–4 students if the students belonged to the same classroom at the same campus. The audit raters strongly agreed (62%) or agreed (38%) that the training and the materials provided during training prepared them well to holistically rate listening and speaking proficiency. The audit raters also strongly agreed (88%) or agreed (12%) that procedural documents provided at audit rater training were clear and complete.

One finding of the pilot was that the time of year of the audit posed logistical challenges for the participants. Audit raters indicated that they would not volunteer again in May because end-of-year duties are too time-consuming. They also indicated that teachers of the students would need to be instructed ahead of time to plan academic lessons that engaged students in more listening and speaking activities. Because of the limited time spent with the students, audit raters sometimes indicated that they would have preferred

to see students in a greater variety of academic listening and speaking interactions to be confident about determining their proficiency levels.

This small-scale study provided the opportunity for TEA to gather feedback from audit raters on logistics and data collection procedures. Input from this pilot will help determine the feasibility of future larger-scale audits and provide information to refine audit and training processes.

Evidence Based on Consequences of Testing

One of the primary reasons the state elected to use a holistic rating component is because of its strong consequential validity. The administration of the TELPAS holistically rated assessments directly supports implementation of the state’s English language proficiency standards (ELPS).

Training of new TELPAS raters initially takes place at the beginning of the school year. The proficiency level descriptors (PLDs), which are the crux of the holistic rating process, are identical to those found in the ELPS. Across content areas, teachers of ELLs are required to teach second language acquisition skills outlined in the standards commensurate with the students’ English proficiency levels as they instruct students in the state’s academic content standards and help them reach higher English proficiency levels. Teachers determine students’ proficiency levels by using the PLDs. The in-depth familiarity that teachers gain with the PLDs as a result of TELPAS rater training helps ensure that the ELPS are, in fact, incorporated in ongoing instruction.

TELPAS thus leads to improvements in students’ academic language acquisition because of what educators learn during the rater training process and through direct application of the PLDs in both ongoing classroom instruction and the live assessment. Shepard (1997) stated that, “a test carefully tied by logical and empirical evidence to the intended content domain is valid for reporting on the status or level of student achievement.” Logical consequences of administering the TELPAS holistically rated assessments are that educators (1) learn how developing academic language proficiency in English relates to and supports academic achievement in English, (2) learn how to interact more with students, adjust content area instruction to meet language needs, and target steady progress in English acquisition, and (3) practice observing student behaviors in the instructional environment for the purpose of making better instructional decisions about students.

Evidence of consequential validity is provided by the positive survey responses gathered from the audits of listening, speaking, and writing. District testing coordinators, campus testing coordinators, and teacher raters from the audited districts and campuses were instructed to complete an audit questionnaire. The results of the questionnaires provide evidence of the efficacy of the training and administration procedures used for TELPAS. The fact that raters volunteered many positive comments about the assessment and that so many comments directly relate to ways in which the assessment is leading to improved teaching and learning are indicators that TELPAS has strong consequential validity.

More evidence of consequential validity has been gathered through an analysis of improved test results. The percentage gains in English language proficiency ratings by domain and for the composite ratings have been consistent and large in the last three school years despite revisions made to the reading tests for grades 2–12. From spring 2006 to spring 2007, gains of from 5 to 7 points were made in the percentage of ELLs in grades 3–12 combined who reached the highest proficiency level (Advanced High) for the composite rating and each separate language domain. From spring 2007 to 2008, gains of from 6 to 8 percentage points were made. In the course of these three years, the percent of ELLs in grades 3–12 combined who received a composite rating of Advanced High rose from 32% to 45%. Finally, substantial gains were seen in ELL performance on the English language arts section of TAKS in secondary grades from spring 2007 to spring 2008. The scores for secondary ELLs on these tests rose on average by 6%.