Scoring TELL (Test of English Language Learning)

By William J. Bonk

Introduction to TELL

Test of English Language Learning (TELL) is an interactive English language proficiency assessment designed to identify, diagnose, and monitor the progress of English language learners (ELLs) in Kindergarten through grade 12 (K-12). TELL allows schools to measure students' English language proficiency at key points during the school year in a flexible, reliable, and consistent way. TELL is delivered automatically on tablet devices, and student responses are scored by Pearson's automated scoring systems that were developed and optimized specifically for TELL assessment tasks and ELL students. Automated delivery and scoring ensures standardization and objectivity in the presentation of test items and in the evaluation of student responses.

TELL presents a series of interactive and engaging tasks such as touching or moving objects with a finger on the screen in response to spoken or written prompts, and watching videos on academic topics and orally summarizing the content. TELL item types are designed to measure the four language skill domains (Listening, Speaking, Reading, and Writing) key to success in school settings. In addition to the four domain scores, some additional subskill scores are also reported in the diagnostic test to provide more granular-level information about a student's English skills. Test scores are usually available within minutes after all test responses for a student have been uploaded.

TELL offers three types of assessments for different purposes: screener, diagnostic, and progress monitor. The tests are specific to groupings of grades that we refer to as grade bands: Kindergarten, 1-2, 3-5, 6-8, and 9-12. Each of the grade bands has its own pool of items and scale. Testing times vary based on the type of assessment administered and the grade band of the

assessment. In general, tests are 20 to 50 minutes in length. A complete package of TELL assessments for one student includes the following tests:

- 1 screener
- 2 diagnostic
- 8 progress monitor

The screener assessment is designed to help determine whether the student qualifies for ELL programs by providing a standardized metric in which the overall English proficiency level of the student is shown. The diagnostic assessments (beginning and end of year) yield detailed information to determine baseline and end-of-year proficiency levels. The progress monitor tests provide relatively frequent longitudinal data to inform instructional decisions over the course of the academic year, and contribute to the construction of an accurate depiction of the trend of a student's proficiency level over time.

Domain and subskill scores

The three TELL assessment types offer different levels of score reporting, as follows:

- screener: overall
- progress monitor: overall, language skill domain scores
- diagnostic: overall, language skill domain scores, subskill scores

Although these score reporting categories are the same for all grade bands, each grade band has its own score scale, as well as domain and subskill scores that reflect the type of assessments administered for that level grade band. The score reporting categories, divided by assessment type and grade band, are summarized in Table 1 below. Possible points on each of the scales are shown in the table; each grade band above kindergarten has a scale that is 100 points higher than the one immediately below it, rather than sharing a common scale. This was done to prevent inappropriate comparisons of scores on different scales; because the scales are not actually linked to one another across grade bands, scores on different grade bands cannot be directly compared.

Table 1. Score Reporting Categories

Grade Band	Score Scale	Screener	Progress Monitor	Diagnostic
K	100 - 200	Overall	Overall Listening Speaking Pre-Literacy Early Writing	Overall Listening* Speaking* Pre-Literacy* Early Writing* Grammar** Vocabulary** Pronunciation** Fluency**
1-2	200 - 300	Overall	Overall Listening Speaking Reading Writing	Overall Listening* Speaking* Reading* Writing* Grammar** Vocabulary** Pronunciation** Fluency** Reading Rate** Pre-Literacy** Reading Comprehension**
3-5	300 - 400	Overall	Overall Listening Speaking Reading Writing	Overall Listening* Speaking* Reading* Writing* Grammar** Vocabulary** Pronunciation** Fluency** Reading Rate** Expressiveness**

Grade Band	Score Scale	Screener	Progress Monitor	Diagnostic
6-8	400 - 500	Overall	Overall Listening Speaking Reading Writing	Overall Listening* Speaking* Reading* Writing* Grammar** Vocabulary** Pronunciation** Fluency** Reading Rate** Expressiveness**
9-12	500 - 600	Overall	Overall Listening Speaking Reading Writing	Overall Listening* Speaking* Reading* Writing* Grammar** Vocabulary** Pronunciation** Fluency** Reading Rate** Expressiveness**

Note. Under Diagnostic, domain scores are marked using *, while subskill scores are marked using **.

The scoring domains and subskills are defined below.

Table 2. Scoring Domains

Overall
The Overall score reflects the student's ability to understand and produce spoken and written English within social and academic communicative contexts at an appropriate grade level.
Listening domain
K-12 The Listening domain score represents the student's ability to comprehend short and extended spoken English of various levels of difficulty from grade-appropriate social language and academic English material.

Speaking domain
K-12 The Speaking domain score reflects the student's ability to produce short and extended spoken English in grade-appropriate social and academic English topics with appropriate vocabulary and grammar.
Reading domain
K The Pre-Literacy domain score reflects the student's ability to demonstrate phonemic awareness, letter-sound correspondence, and knowledge of print concepts.
1-2 The Reading domain score reflects the student's ability to demonstrate knowledge of print concepts and to accurately read and understand word and sentence level text in grade-appropriate social language and academic English topics.
3-12 The Reading domain score reflects the student's ability to comprehend written English texts. These texts vary in their readability, length, and content (i.e., social versus academic language). Comprehension and reading skill are demonstrated by smoothly and accurately reading written texts out loud, and by various tasks that assess comprehension of the texts' content.
Writing domain
K The Early Writing domain score reflects the student's ability to copy and write letters and words in English.
1-2 The Writing domain score reflects the student's ability to write words and phrases in English in grade-appropriate social language and academic English topics with appropriate vocabulary and grammar.
3-12 The Writing domain score reflects the student's ability to write and organize words and sentences in English in grade-appropriate social language and academic English topics with appropriate vocabulary and grammar.

Table 3. Subskills

Fluency
Fluency subscores are given based on the student's speaking ability. This score reflects the rhythm, phrasing, and pausing used during a continuous flow of speech.
Pronunciation
Pronunciation subscores reflect the accuracy of vowel and consonant pronunciation, phonological form, and stress placement.
Grammar
Grammar subscores reflect how accurately and coherently sentences are combined to thoroughly convey main ideas and relevant details. Level of control means using conventional English effectively and accurately.
Vocabulary
Vocabulary subscores represent a student's ability to understand general and academic words in spoken and written sentence contexts and to produce such words as needed. Performance depends on familiarity with the form and meaning of general and academic words and their use in connected texts.
Reading rate
Reading Rate subscores reflect the number of words correctly read per minute during the oral reading task.
Expressiveness
Expressiveness subscores reflect the student's ability to derive meaning from text while reading aloud, as demonstrated by using correct stress patterns, phrasing, and pausing to convey meaning.

Reading comprehension
Reading comprehension subscores reflect the student's ability to understand written text and correctly respond to given commands.
Pre-Literacy
Pre-literacy subscores are given based on the student's emerging literacy skills in grades 1-2. This score reflects understanding of sound-symbol correspondence, written English conventions, and word-level decoding abilities.

Score aggregation

As part of the test development process, a large-scale field test was conducted between January and March 2015 to collect data from both ELLs and from students identified as English native speakers. Over 10,000 tests were completed during field testing, from a roughly equal number of students in the five grade bands. In total, 71 schools from 23 U.S. states, one U.S. territory, and American schools in Brazil and Mexico participated in the study. The student participants used a prototype version of TELL to take the field tests on iPads.

A primary goal of this field testing was to collect responses for the entire pool of test items from a large sample of English language learners at various levels and with various first language backgrounds in order to launch TELL with a robust item calibration. As described in the TELL Technical Manual (Pearson, 2015), Rasch models (Rasch, 1960/1980) were constructed for each domain score (Reading, Writing, Listening, and Speaking), as well as for Pronunciation and Fluency scores. This process of transforming the test-taker responses scored individually into language domain skill scores is called psychometric modeling. Psychometric modeling enables test developers to put all items in a domain onto a single scale (item calibration), which provides both a useful item analysis during development, and a scoring method after development is complete. Some of the output of this modeling provides information on the quality of each item; items of lower quality are identified and excluded from operational tests. This field test psychometric modeling also forms the basis for scoring tests once the test becomes operational – this modern method of score generation contrasts with the more familiar "add up all the points earned, divide by the total points possible, and give the score as a percentage correct" method.

When a student takes TELL, a set of scores is generated in the areas listed in Table 1 (depending on the type of test taken). Overall scores are an equally weighted combination of Listening,

Speaking, Reading, and Writing. Subskill scores listed for Diagnostic tests are separately calculated and reported, but do not feed into Overall scores, because those item responses are already included in the domain scores. In other words, subskill scores are simply a more granular view of the domain scores.

Pearson's automated scoring systems

A number of systems owned and operated by Pearson have the capability to score student responses automatically. The response scoring system at the heart of TELL was developed by leveraging the knowledge from two decades of experience in speech recognition, natural language processing, machine learning, and assessment construction. These scoring systems already operate in a number of Pearson products and services such as WriteToLearn, Versant language tests, PTE-Academic, and PARCC, among many others.

These automated scoring systems are at the deepest level based on human judgment. First, test developers decide what aspects of language should be measured, and design item types to elicit them. Then they write meaningful scoring rubrics – the ratings using these are to be fed into the machine learning models. To prepare the input to the automated scoring models for spoken portions of a test, a team of expert linguists is trained to transcribe a large number of responses. Transcriptions (rather than sound files) constitute the input for modeling that is focused on language use or content. These transcribed responses are evaluated by human raters for content and language use using the scoring rubrics so that their scoring is based only on *what* was said rather than on *how* it was said. In this manner, test-takers' pronunciation, fluency, and other qualities of the way they sound do not have an influence on content or language use ratings. Those qualities of the manner of speaking are separately assessed by raters listening to the responses and rating them with other scoring rubrics on traits such as pronunciation and fluency.

Trained raters use scoring rubrics to evaluate a large number of real test-taker performances, and the quality of those sets of ratings is assessed. If the ratings are reliable enough (i.e., if the same performances tend to get the same or similar ratings from a number of independent raters), they can be used to train the automated scoring system. An example of one of the rubrics used is shown in Figure 1. This rubric was used by raters to evaluate the content of the Listen and Retell item type, in which students heard an extended narrative, and were asked to retell it in as much detail as they could. Note that the scoring rubric does not require students to use the same words presented in the narrative. The focus is on whether or not they accurately reproduced the ideas contained in it.

Rating		Descriptors
3	Accurate performance	Most or all important ideas from the text accurately represented. Little or no content is inaccurately represented.
2	Somewhat accurate performance	Some important ideas from the text accurately represented. Some ideas from the text are missing or inaccurately represented.
1	Limited accuracy	Few or no important ideas from the text accurately represented. Many ideas from the text are missing or inaccurately represented.
0	Not scorable	Response meets one or more of the following conditions: • Unintelligible text • Response is off-topic • Response is 5 words or less

Figure 1. TELL "Listen and retell" item type content scoring rubric

The quality of the machine scoring depends to a great extent on the quality of the ratings that go into the models as training data. If expert human raters disagree on the scores to assign to particular performances, then that confusion will in turn become a part of the machine scoring algorithm. On the other hand, if the expert raters generally agree on the correct rating for particular responses, then the machine scoring system will learn how to score other responses with similar characteristics, and it will result in a reliable system. In other words, the automated system acts like a human rater when assessing test takers' language skills but does so with the precision, consistency, and objectivity of a machine. There is evidence that Pearson's automated scoring can outperform human graders in consistency (Foltz, Streeter, Lochbaum, & Landauer, 2013).

During field testing, expert raters were contracted by Pearson to rate a number of responses using the scoring rubrics. Typically, several hundred responses are rated for each item. Those raters are first given training sessions with actual student responses. They practice using scoring rubrics like the one shown in Figure 1. After a norming session allowing for discussion and feedback on where they seemed to be both on- and off-target with their practice ratings, raters are given a test set to rate. If their ratings are sufficiently accurate for the test set, they "pass" and are contracted to participate in the actual rating sessions. Raters who are not successful in producing a correct set of ratings on the test set are not used as raters. Those who pass will read or listen to responses independently (i.e., without seeking consensus with other raters), and

provide ratings in an online tool. The responses are presented to raters in a randomized order so that there is no discernible pattern to influence their ratings. The data are all collected in a database until the desired number of ratings for each response (depending on the score domain, typically 2-3) is reached.

Once that process is complete, the ratings are examined for inter-rater reliability and for quality indicators such as sufficient numbers of responses in the various rating categories. If the dataset is found to be adequate, it will be modeled. Modeling typically involves using some already built underlying infrastructure such as a Latent Semantic Analysis model (Landauer, Laham, & Foltz, 2003) to analyze the dataset. In this case the dataset is a few hundred responses to a particular test item, along with the human ratings associated with those responses. A scoring model is built for each item – the model uses characteristics of what is in the responses (such as particular combinations of words), consults the expert ratings, and in an iterative fashion "learns" how the human raters seemed to react to the many characteristics of the responses. It learns that humans gave high ratings to some kinds of characteristics, and low ratings to others. For example, in a written response that should be a summary of a passage that the test-taker read, expert raters would tend to give higher scores to responses that include words and phrases that have the same meaning as what was in the original passage, and have grammar and spelling that conforms to the expectations of English. In the end, the machine learning algorithm settles on the best possible solution for that dataset - a solution that effectively reproduces the human ratings on which it was originally trained. Essentially, it has "learned" to be sensitive to the same kinds of features in responses that expert raters were sensitive to, so it imitates their behavior and rates responses the same way that they would.

A model developed in this way may sometimes get particularly good at reproducing the ratings that it was fed during its development, so a rigorous check on its quality is separately performed. How well does the model approximate expert ratings for responses on which it was not trained? This process is called machine scoring validation. A large number of new, unseen responses are fed into the system, and the model generates its machine scores. Those scores are compared to the scores independently provided by expert human raters for each item. If there is a close correspondence between the machine scores and the ratings provided by experts, then the model built for that item is considered successful. If the machine and human scores differ to a great extent, that item will not be included when the test becomes operational because the scoring model developed for it was rejected.

TELL item types and traits

TELL item types

As shown by the many different types of scores produced, TELL measures many aspects of English proficiency. In order to do this well, TELL utilizes a number of different item types (shown in Table 4) designed to elicit responses that can serve as the solid basis for determinations of proficiency.

For example, the item type called "Listen and act" presents test takers with a number of images on a background scene on their touch tablet. Test takers hear an instruction such as "Touch the object that was invented in the 20th century" and have 15 seconds to carry out the action on screen using their finger. These instructions are carefully constructed so that they contain language features associated with target English language development standards (e.g., those of WIDA (2012). Some of these features include grammatical complexity, vocabulary, and academic language. These items include pictures that serve as distractors so that students cannot simply guess the correct answer. In contrast to many other tests of listening comprehension, this listen and act task can be said to tap into listening comprehension rather directly. More traditional listening items often present a listening passage, followed by a multiple choice question that must be read. This kind of item requires both reading and listening ability, whereas the TELL item isolates listening comprehension in a purer fashion.

All the TELL item types shown in Table 4 went through a similar decision process during the test design phase. Some item types isolate particular skills, while others integrate skills that fit together well, such as reading, then writing a response in the "Read and summarize" item type.

Table 4. Summary of TELL item types

Item Type Name	Item Type Description	Skills Required	Grade Band	Response Type	Question Type	Traits scored
Say the word	Verbally identify image shown on screen (e.g., "Say what's in the picture.")	Listen-Speak	K	Short spoken response	Performance task	• Accuracy in producing the key word(s)
Pick the right picture	Identify by touch, the desired basic text feature from three images (e.g. "Touch the page with the word X.")	Read/Print Awareness	K	Selected response	Technology-enabled	• Accuracy in selecting the correct answer choice
Say the letter	Read list of 5 upper- and lower-case letters aloud (e.g. "Now read these letters out loud.")	Read/Early Literacy	K	Short response	Performance task	• Accuracy in saying the correct sounds in the correct order
Copy the letter	Copy (with finger) the letter displayed on the screen (e.g. "Write the letter X."	Write/Early Literacy	K	Short response	Technology-enhanced	• Accuracy in producing the correct letter shapes on the screen
Copy the word	Copy (with finger) the word displayed on the screen (e.g. "Write the word X.")	Write/Early Literacy	K, 1-2	Short response	Technology-enhanced	• Accuracy in producing the correct letter shapes on the screen

Item Type Name	Item Type Description	Skills Required	Grade Band	Response Type	Question Type	Traits scored
Pick the right letter	From a group of 3 letters, touch the one that represents the sound played (e.g. "Touch the letter that makes the sound X.")	Read/Print Literacy	K	Selected response	Technology-enabled	• Accuracy in selecting the correct answer choice
Pick the right word	From a group of 3 letters or 3 words, touch the one that represents the sound played (e.g. "Touch the word that starts with the sound X.")	Read/Print Literacy	1-2	Selected response	Technology-enabled	• Accuracy in selecting the correct answer choice
Find the error	Identify the word with a spelling or capitalization error (e.g. "Touch the word that's wrong.")	Read/Print concepts	1-2	Selected response	Technology-enabled	• Accuracy in selecting the correct answer choice
Write the word	Handwrite the word that is heard with corresponding image (e.g. "This is a X. Now you write the word X.")	Write	1-2	Short written response	Performance task	• Accuracy in producing the correct letter shapes on the screen
Write about the picture	Write a description for the picture shown on screen (e.g. "Write about what's happening in the picture.")	Write	1-2	Extended written constructed response	Open-ended, Performance task	• Evaluation of the response content based on a scoring rubric

Item Type Name	Item Type Description	Skills Required	Grade Band	Response Type	Question Type	Traits scored
Read the words	Read a list of words aloud (e.g. "Read these words out loud.")	Read-Speak	1-2	Short response	Technology-enhanced	• Accurate spoken production of the words presented
Describe the video	Watch silent video and describe its content in complete sentences	Speak	1-2	Spoken constructed response	Open-ended, Performance task	• Evaluation of the response content based on a scoring rubric
Listen and act	Listen to a prompt and then touch or move a specified object (e.g. "Touch/move the X.")	Listen	All grade bands	Selected response	Technology-enhanced	• Accurate production of the target action in the target location(s)
Repeat the sentence	Listen to a short sentence and repeat it verbatim	Listen-Speak	All grade bands	Short response	Performance task	• Accurate reproduction of the target words in the correct order • Evaluation of the response pronunciation based on a scoring rubric • Evaluation of the response fluency based on a scoring rubric

Item Type Name	Item Type Description	Skills Required	Grade Band	Response Type	Question Type	Traits scored
Listen and retell	Listen to an audio passage and retell it in your own words	Listen-Speak	All grade bands	Extended spoken constructed response	Performance task	• Evaluation of the response content based on a scoring rubric • Evaluation of the response language use based on a scoring rubric • Evaluation of the response pronunciation based on a scoring rubric • Evaluation of the response fluency based on a scoring rubric
Read and act	Read a prompt on screen and then touch or move a specified object	Read	1-2, 3-5, 6-8, 9-12	Selected response	Technology-enhanced	• Accurate production of the target action in the target location(s)
Complete the sentence	Read a sentence with a missing word and then type the word in the blank	Read-Write	1-2, 3-5, 6-8, 9-12	Short response	Technology-enabled	• Accurate production of an appropriate word for the sentence

Item Type Name	Item Type Description	Skills Required	Grade Band	Response Type	Question Type	Traits scored
Put the word in the sentence	Drag the word on top of the screen to make a grammatically correct sentence	Read	3-5, 6-8, 9-12	Selected response	Technology-enhanced	• Accurate placement of the target word into the sentence
Speak in the situation	Listen to an audio prompt describing a situation and respond appropriately in speaking	Listen-Speak	3-5, 6-8, 9-12	Spoken constructed response	Performance task	• Evaluation of the response's situational appropriateness based on a scoring rubric • Evaluation of the response pronunciation based on a scoring rubric

Item Type Name	Item Type Description	Skills Required	Grade Band	Response Type	Question Type	Traits scored
Watch and explain	Watch a video of a mini lecture explaining a concept and then summarize the content by speaking, followed by a comprehension question	Listen-Speak	3-5, 6-8, 9-12	Summary: Extended spoken constructed response Comprehension Question: Short response	Summary: Open-ended, Performance task Comprehension Question: Performance task	• Evaluation of the response content based on a scoring rubric • Evaluation of the response language use based on a scoring rubric • Evaluation of the response pronunciation based on a scoring rubric • Evaluation of the response fluency based on a scoring rubric • Accuracy in producing the key word(s) for the comprehension question

Item Type Name	Item Type Description	Skills Required	Grade Band	Response Type	Question Type	Traits scored
Read the passage	Read a passage aloud and answer a comprehension question verbally	Read-Speak	3-5, 6-8, 9-12	Oral reading: Extended spoken constructed response Comprehension Question: Short response	Oral reading: Performance task Comprehension Question: Performance task	• Percentage accuracy in spoken production of the words presented • Evaluation of reading expressiveness based on a scoring rubric • Accuracy in producing the key word(s) for the comprehension question • Evaluation of the response pronunciation based on a scoring rubric • Words correct per minute rate

Item Type Name	Item Type Description	Skills Required	Grade Band	Response Type	Question Type	Traits scored
Read and summarize	Read a text passage and then summarize it in writing, followed by a comprehension question	Read-Write	3-5, 6-8, 9-12	Summary: Extended constructed written response Comprehension Question: Short response	Summary: Open-ended, Performance task Comprehension Question: Performance task	• Evaluation of the response content based on a scoring rubric • Evaluation of the response language use based on a scoring rubric • Accuracy in producing the key word(s) for the comprehension question

Scoring for TELL item types

As shown in Table 4, TELL has 22 item types, each with its own unique design, content, and trait(s) to be scored. The column in this table labeled "Traits scored" includes a list of all the traits, or aspects of the response that are automatically scored. Scoring methods for these traits depend on the response method and what the traits are, and can be roughly grouped as follows.

Selected responses. Test takers need to select the correct item on screen within the time limit to get credit for a correct response. Otherwise, their response is considered incorrect. If a location or action is specified, then these are also required for a response to be considered correct.

Short responses. These responses require the production of only a small number of letters or words. In the case of letters or words written with the finger on the tablet, a handwriting recognition model operates in the background to determine if the correct letters or word were written, and it scores the response as correct or incorrect. In the case of a spoken or written word or short phrase, the scoring system consults a dictionary of correct keywords; matches between the entries in that dictionary and a response are scored as correct. The system is flexible enough to allow spelling errors or somewhat incorrectly pronounced words as matches where appropriate. In the case of the "Repeat the sentence" item type, where an exact repetition of the sentence is requested, partial credit is given when some but not all of the target words are spoken by a test-taker.

Extended constructed responses. The scoring system uses a number of advanced models like LSA (Landauer, Laham, & Foltz, 2003) to score the content of these responses. Some aspects of the process of building these models have already been described above. Responses are scored by the model specifically built for that prompt as would a human rater – the system assigns a "partial credit" score as if it were reading the response and consulting the scoring rubric. This is a partial credit score because responses can have a range of scores, not just correct/incorrect, as would be the case for a multiple choice item. Machine scores are even finer-grained than values on the scoring rubric because they have decimal places, but those values are rounded to the nearest integer in score production.

In the case of pronunciation, fluency, and expressiveness, existing Pearson models for those traits were used for TELL. The existing models did not need to be trained on TELL data because they are not item-specific; they analyze the speech stream and calculate values for certain key features (such as inter-word silence) that appear in responses regardless of the content. A complete

discussion of what sorts of features they analyze and how they work is beyond the scope of this document, but a summary can be found in Bernstein, Van Moere, and Cheng (2010) (available upon request from Pearson).

In the case of reading aloud, the percentage of the words accurately read and recognized by the system is calculated. Although this is an extended constructed response, an advanced model is not required because the meaning of the content is not analyzed, only the words spoken. This value is also used to calculate a measure of reading rate: words correct per minute. This reading rate is adjusted for accuracy so that test takers do not get credit if they do not actually read the material they should be reading, or for saying only certain words over and over again quickly in an effort to "game" the test.

Reliability and Validity

Test reliability

The reliability of a test refers to the consistency of test scores – did the student receive the score that he/she deserved? When reliability is high, students of similar ability would get similar test scores because measurement error is held to a minimum.

Measuring the reliability of scores can be approached in various ways. One type of reliability estimate is known as *alternate forms test-retest reliability*. In this approach, parallel tests, with different sets of items, are administered to the same students on more than one occasion, with the assumption that students' underlying proficiency level did not change between these occasions, and that the students' test scores on both those occasions can be considered independent. If there is a high correlation between the pairs of scores for the same participants, reliability is high. If there is greater variation from one testing occasion to the next in test scores for the same participants, there is a greater amount of measurement error, and reliability is low.

In TELL field testing, 1,226 students successfully took the prototype version of TELL two or more times within the same week. Scores for those students on each test they took were calculated once the final psychometric modeling and scaling were completed, and then matched up. Their test forms were randomly assembled from a larger pool of items so that the student did not receive the same set of items on different tests. Correlations were calculated for those matched sets of test scores, shown in Table 5. All correlations were positive and statistically significant. Most correlations showed good reliability for TELL, especially for students in grades 3 and above when they are better at following instructions and speaking and writing clearly in general.

Table 5. Alternate forms test-retest reliability for overall scores and four language skill domain scores

Grade Band	*n*	Overall	Listening	Speaking	Reading	Writing
K	229	.70	.72	.69	.44	.32
1-2	332	.79	.55	.58	.85	.74
3-5	383	.91	.84	.85	.79	.79
6-8	127	.87	.80	.77	.78	.70
9-12	155	.83	.73	.73	.78	.63

Another way that test reliability can be estimated is with internal consistency estimates. If a test's items are internally consistent, then a test taker would get similar scores using different arbitrary groupings of those items. Often this is directly calculated using the split-half method (Brown, 1996). For example, if a student took a 60-item test, we split the test into two halves, A and B, and calculate the student's score on the 30 form A items and on the 30 form B items separately. If scores A and B are very different, reliability is low, because those collections of items from the same test gave different results. Conversely, if the scores are the same or very close, this indicates high reliability. A more robust form of this calculation is Cronbach's alpha (Cronbach, 1951), which is essentially the average of all possible split-half reliability coefficients. These range from zero to 1, with values closer to 1 indicating higher test reliability.

During psychometric analysis, an internal consistency estimate akin to Cronbach's alpha was also calculated for each grade band and domain score as they were modeled in Winsteps (Linacre, 2015). The obtained estimates are shown below in Table 6. Most values show good or excellent reliability.

Table 6. Internal consistency (reliability) estimates for four language skill domain scores from psychometric software calculations

Grade Band	Listening	Speaking	Reading	Writing
K	.76	.82	.51	N/A
1-2	.74	.83	.88	.66
3-5	.85	.92	.87	.72
6-8	.88	.94	.86	.79
9-12	.91	.95	.88	.79

Reliability estimates for Kindergarten Reading were not as high as the others. This is likely due to the same reason as for Grades K-1-2 Writing. That is, because TELL Kindergarten and to some extent grades 1-2 Writing items only tapped into a very basic level of writing proficiency, discrimination is mainly at the lower levels for these scores, and overall reliability estimates either appear lower than would normally be required, or cannot even be calculated using the typical method (as in the case of Kindergarten Writing here) because of inadequate variance in the data.

The overall pattern of the results of these two analyses converge to demonstrate TELL's highly reliable item pool and scoring system. Because of a successful concurrent item calibration, different subsets of items presented to the same people yield very similar scores on both occasions, and the estimates of internal consistency show high test score reliability on the various language skill domains.

Inter-rater reliability

As previously described, multiple independent raters scored field test responses using scoring rubrics in the development of TELL. The machine learning algorithms use averages of those ratings as a "true" rating for a test taker response; those "true" ratings are what the machine learns from. A different measure of reliability pertinent to TELL is how well those human raters agreed with one another. Table 7 shows the correlation coefficients for pairs of expert human ratings for a number of item types and traits, those yielding the longest responses on TELL and which are presented to the widest range of grade bands. These responses were a part of the holdout dataset set aside at the conclusion of data collection (described in detail in the next section of this document). In other words, these ratings were not used as a part of the dataset used to train the machine learning system.

Table 7. Inter-rater reliability correlations for TELL extended spoken and written response ratings

Item Type	Trait scored	n-count of responses	Inter-rater correlation
Listen and retell	Content	2,100	.85
	Language use	2,100	.78
Watch and explain	Content	981	.80
	Language use	981	.76
Read and summarize	Content	672	.70
	Language use	672	.59
Speak in the situation	Appropriateness	1,800	.88

Higher correlations indicate higher reliability, or agreement among raters. As shown in Table 7, the inter-rater correlations are in most cases quite high, consistent with data from other operational tests, thereby creating a solid base from which the machine learning algorithms can operate. It should be noted that the data in Table 7 are from the field test prior to the finalization of TELL. The data contain ratings on a large number of items that did not have sufficiently high enough quality to be included in the operational test. As a result of the analysis of field test inter-rater reliability values, machine-human correlations, and psychometric analysis, many items are excluded. This table presents the results *prior to* these exclusions, so the values in it represent the lower bounds of the true inter-rater reliability for items actually included in the operational TELL.

Machine-human-score correlations

A key analysis of a test with automated scoring is a determination of the accuracy of machine scores compared with human judgments; in particular, the extent to which TELL scores accurately reflect the scores that human raters would assign to the test takers' spoken and written performances.

In order to address this question, a sample of field-test participants from each grade band was kept separate from the main dataset; this serves as holdout dataset for the purposes of later validating the machine scoring model once it has been developed. In each grade band 150 field test-takers (750 total) were randomly assigned to the holdout dataset at the conclusion of field testing. The only constraints on selection into this condition were: (1) a successfully completed test session, (2) approximately 95% of the test takers were identified as English Language Learners by

the school participating in field testing, and (3) approximately 5% were reported to be native speakers of English.

These 750 students, or the teachers who helped organize test administrations during field testing, were not aware that their data would be treated differently, nor were the tests administered in any way differently. These test takers were not identified for inclusion in the holdout dataset until after all tests had already been administered. Responses from those test takers were carefully sequestered off so that they were not included either in the item calibration phase of psychometric analysis or in the training and development phase of automated scoring systems. Machine training consists of both automatic speech processing and the machine learning of how to assign ratings on the basis of features in the responses and expert human ratings.

The responses for the 750 test takers in the holdout dataset were submitted for expert rating and transcription to produce human-based scores. This process of collecting expert ratings and transcriptions followed the same process previously described for developing the training set for the machine learning system to use, but the ratings are used for a different purpose: to produce human-based scores instead of machine-based scores. For example, a Listen and Retell response would be transcribed by trained transcribers and then, using a set of scoring rubrics, human raters would rate the transcriptions for content accuracy and language use, and the recordings of responses for fluency and pronunciation; a Read and Summarize response would be scored by human raters for content accuracy and language use; or a Copy the Letter item would be assigned a correct/incorrect score by a trained rater. Some item types did not lend themselves to this treatment because human judgments were unnecessary. For example, a multiple choice test question has a previously determined correct answer, and further human judgment is not required if a student's response was correct. Therefore, human ratings or transcriptions were used wherever they were available to create human ratings-based scores for the purposes of this comparison. When they were not available, the objectively scored response was used in both machine-based and human-based scores.

The goal of this validation exercise is to determine if test takers' scores are comparable when calculated by either the machine scoring method or from human ratings. To perform this analysis, machine-generated scores and human-based scores were produced for each test taker in the holdout dataset for each language skill domain, as they would be on the operational test, including all items presented. This comparison can be considered a valid and realistic way to determine how closely machine-generated scores mirror human-based scores of the same test performances because it exactly replicates the operational scoring method and test length, substituting in human scores wherever they are available.

In TELL, domain scores in Reading, Writing, Listening, and Speaking are calculated using a Rasch model. These Rasch model scores are then transformed into TELL scale values, truncated at the upper and lower limits of each scale where necessary, and rounded to integers. Overall scores are the rounded average of the four domain scores. In some cases, score estimates could not be generated for particular test takers in a given language skill domain. This occurred when, for example, many items on a test taker's form were shown not to have adequate psychometric qualities for inclusion in the operational test, so the items were "killed," and consequently those items were not included in the scoring system and that test taker did not have enough valid items from which to create a score. When a test taker was missing one of the four language skill domain scores, his/her average for that scoring method (human- or machine-based) was not calculated, and the case was excluded from the analysis. Table 8 shows the results of this analysis. The closer a correlation of machine to human scores approaches 1, the more closely the rank order of scores of each set of 150 test takers is the same, regardless of scoring method.

Table 8. Correlations of human- and machine-based scoring methods by grade band

Grade Band	Overall	Listening	Reading	Speaking	Writing
K	.78	.89	.99	.70	.38
1-2	.88	.88	.79	.79	.96
3-5	.92	.91	.90	.85	.96
6-8	.90	.86	.91	.77	.96
9-12	.83	.76	.91	.59	.97

The analysis shown in Table 8 demonstrates that TELL automated scoring produces test scores that are, overall, highly correlated with scores derived mostly from human judgment. These are obviously only estimates; actual operational test reliability may differ somewhat because of the quality of the field test data and the match of the field-test participants to the tasks presented. During scoring, the automated speech recognition system has internal checks on its own performance, a "confidence in recognition" metric. When a great deal of background noise is present, or when test takers do not speak clearly or intelligibly enough, the system's confidence level tends to be low. Low confidence levels generate internal alerts that would prevent a spoken response-based language skill domain score from being released and aggregated with other scores. These confidence level alerts were frequently generated among grade 9-12 test takers, decreasing the reliability of automated scoring. Removing the cases where this alert was generated substantially increased the machine-human correlations for these students: the

observed correlations were .95, .89, and .83 for overall, listening, and speaking scores respectively when low confidence scores were not included. The listening and speaking scores for those students with unintelligible content are suppressed during operational testing, so actual machine scoring reliability is likely to be somewhat higher than that reported in Table 8.

Discussion

TELL presents an elegant solution to the problem of how to quickly and accurately measure K-12 students' English language development. TELL is made up of a number of technology-enhanced item types, many of which require students to produce constructed responses in English. Because it requires productive as well as receptive language skills, TELL is able to calculate test scores on the basis of actual language use, not just knowledge *about* language, or ability to get correct answers on multiple choice questions. These novel item types are designed to elicit responses with characteristics that are key indicators of language proficiency, thus enhancing the authenticity of the tasks, and supporting the validity of inferences made from the test scores. TELL language skill domain and subskill scores give a detailed profile of student proficiency in English, identifying areas of both strength and weakness, as well as a summary overall proficiency level.

The sophisticated automated scoring techniques used in TELL make it possible to generate these scores quickly, and without the need for teacher or other human rater time. In this document, we have described the process of how we develop and test automated scoring for the TELL item types: the specific design of the TELL items types and the features in them to be scored; the data collection for item calibration and model building; the transcription, human rating, and machine training phase; the psychometric modeling and quality control processes around the promotion of items to the operational item bank; and finally, the robust analyses of reliability and of the generalizability of the results of the machine scoring algorithm to new data. As a result of thoughtful, data-driven, and validated construction, the resulting automated scoring system provides consistent, objective test scores in a fashion timely enough to be useful in guiding classroom decisions and evaluating the ongoing effectiveness of English language services.

References

Bernstein, J., van Moere, A., & Cheng, J. (2010). Validating automated speaking tests. *Language Testing, 27*, 355-377.

Brown, J. D. (1996). *Testing in language programs.* Upper Saddle River, NJ: Prentice Hall Regents.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297–334.

Foltz, P. W., Streeter, L. A., Lochbaum, K. E., & Landauer, T. K. (2013). Implementation and applications of the Intelligent Essay Assessor. In M. Shermis & J. Burstein (Eds.), *Handbook of Automated Essay Evaluation* (pp. 68-88). New York, NY: Routledge.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to Latent Semantic Analysis. *Discourse Processes, 25,* 259-284.

Linacre, J. M. (2015). *Winsteps*® (Version 3.90) [Computer Software]. Beaverton, OR: Winsteps.com.

Pearson (2015). *Test of English language learning technical manual*. Menlo Park, CA: Pearson. *(*Available upon request from publisher.)

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. (Copenhagen, Danish Institute for Educational Research). Expanded edition (1980) with foreword and afterword by B.D. Wright. Chicago: The University of Chicago Press.

World-Class Instructional Design and Assessment (WIDA) (2012). *Amplification of the English Language Development (ELD) Standards*. Retrieved from http://www.wida.us/standards/eld.aspx