Pearson's Automated Scoring of Writing, Speaking, and Mathematics

White Paper

Lynn Streeter, Ph.D.
Jared Bernstein, Ph.D.
Peter Foltz, Ph.D.
Donald DeLand

May 2011

About Pearson
Pearson, the global leader in education and education technology, provides innovative print and digital education materials for preK through college, student information systems and learning management systems, teacher licensure testing, teacher professional development, career certification programs, and testing and assessment products that set the standard for the industry. Pearson's other primary businesses include the Financial Times Group and the Penguin Group. For more information about the Assessment & Information group of Pearson, visit http://www.pearsonassessments.com/.

About Pearson's White Papers
Pearson's white paper series shares the perspectives of our assessment experts on selected topics of interest to educators, researchers, policy makers and other stakeholders involved in assessment and instruction. Pearson's publications in .pdf format may be obtained at: http://www.pearsonassessments.com/research.

Executive Summary

The new assessment systems designed to measure the Common Core standards will include far more performance-based items and tasks than most of today's assessments. In order to provide feedback quickly and make scoring of the responses affordable, the new assessments will rely heavily on artificial intelligence scoring, hereafter referred to as automated scoring. Familiarity with the current range of applications and the operational accuracy of automatic scoring technology can help provide an understanding of which item types can be scored automatically in the near term. This document describes several examples of current item types that Pearson has designed and fielded successfully with automatic scoring. The item examples are presented along with operational reliability and accuracy figures, as well as information of the nature and development of the automated scoring systems used by Pearson.

In the 1990s, the people of Pearson's Knowledge Technologies group (KT) invented many of the key techniques that enable automatic scoring of constructed language in assessment tasks. In the succeeding 15 years, Pearson has assembled these researchers into an advanced development group with an intellectual property base that is unparalleled in the assessment field. Now, working as a unique stand-alone group inside Pearson, KT has automatically scored many millions of written and spoken responses. KT has measured core language and literacy skills as evidenced in students' constructed responses. Similar tasks also elicit responses that are assessed for content knowledge. In 2010 KT scored over 20 million spoken and written responses from all over the globe.

The document introduces a framework for understanding Pearson's work in automated scoring. KT automatically scores written, spoken, and to a lesser extent mathematical responses. For most aspects of writing and speaking, the performance of automated scoring already equals or surpasses that of human raters. Written text can be scored for language arts and content domains, such as science, social studies, and history. Written texts are scored for declarative knowledge and language skills as reflected in stylistic and mechanical aspects of writing. Spoken responses are scored for declarative knowledge and speech quality in tasks such as reading aloud to determine fluency and in orally summarizing a reading. The most extensive applications of spoken technology are determining proficiency in speaking and understanding English as well as other languages. Written and spoken automated scoring have been combined to assess the traditional four language skills (reading, writing, speaking, and listening) for college admission and employment decisions. Pearson's automated mathematics assessments are under development and will allow students to show and explain their work as they step through computations, derivations, and proofs using graphic and equation editors and text input.

KT has scored constructed responses for use in primary, secondary and post-secondary education, as well as for governments, publishers, and other corporations. Pearson's accumulated experience with large-scale automatic scoring affirms the feasibility of using our technology in emerging assessments, such as those envisioned in the Common Core State Standards (CCSS). These new assessments will require complex, authentic task types that elicit integrative student performances and can be scored automatically with available technologies.

Examples of items and actual assessments are described throughout the paper along with comparisons with human judgments where available. The examples of automated scoring technology are intended to inform educators, policy makers, and test developers about the state of the field and its potential for use in near-term and future assessment systems.

Pearson's Automated Scoring of Writing, Speaking, and Mathematics

Introduction

The Common Core State Standards (CCSS) advocate that students be taught 21st century skills, using authentic tasks and assessments. Students should exhibit effective content knowledge and independent thinking in their ability to critique information, evaluate evidence, and make sense of problems in order to solve them. New assessments will incorporate more items that require students to demonstrate their problem solving skills on authentic, complex tasks in language arts, mathematics, and other content areas. The use of constructed-response (CR) items will grow, increasing reliance on human scoring, intelligent computer scoring or a combination of both.

Automated scoring systems provide consistency over time and location, which promotes equity, enables accurate trend analysis, and provides comparable results for use at the classroom, school, district, or state level. Automated scoring of CR items has grown rapidly in large scale testing because systems can produce scores more reliably and quickly and at a lower cost than human scoring (see Topol, Olson, & Roeber, 2011). There are several automated systems (at Pearson and elsewhere) able to score CR items, including essays, spoken responses, short text answers to content questions, and numeric and graphic responses to math questions. Generally, scoring systems for CR items require digital delivery of items and entry of responses. As with human scoring, the accuracy of automated scoring depends on several factors, including task clarity and well-designed training data. For automated scoring systems, the degree of constraint expected in the constructed responses is somewhat more important than it is with human ratings.

Pearson's Knowledge Technologies

In the 1990s, the people of Pearson's Knowledge Technologies group (KT) invented many of the key techniques that enable automatic scoring of constructed language in assessment tasks. In the succeeding 15 years, Pearson has assembled these researchers into an advanced development group with an intellectual property base that is unparalleled in the assessment field. Now, working as a unique stand-alone group inside Pearson, KT has automatically scored many millions of written and spoken responses. KT has measured core language and literacy skills as evidenced in students' constructed responses. Similar tasks also elicit responses that are assessed for content knowledge. (Additional knowledge about products and research can be found at www.pearsonkt.net; www.versant.com; writetolearn.net.) In 2010 KT scored over 20 million spoken and written responses from all over the globe.

Evaluating Automated Scoring Systems

How does one know automated scoring works well enough for a high-stakes test? The following two questions (adapted from several in Williamson et al., 2010) are useful for evaluating any automatic scoring system.

1. *Are automated scoring methods disclosed in sufficient detail to determine their merit?*
2. *Are automated scores consistent with consensus scores from human graders?*

This document focuses on these questions in relation to currently operational CR item types that Pearson scores automatically. The document provides examples of item types now being scored with descriptions of the scoring methods. Much more detailed information on item types and how they are scored is available in other published technical articles and in Pearson technical reports (e.g. Bernstein & Cheng, 2008, Foltz et al., 2000, Landauer et al., 2002). The most basic question is: does an automatic system produce consistently accurate scores? This is usually answered by addressing question 2, above. Table 1 presents correlation coefficients between automated scores and consensus human scores for a sample of written and spoken constructed responses.

Autoscoring Performance					
Response	**Assessment Prompt Material**	**N**	**Machine-Human** Score Correlation	**Human-Human** Score Correlation	**Source**
Written	81 published essay prompts (grade 6-12)	400	0.89	0.86	Prentice Hall
	18 research-leveled essay prompts (grade 4-12)	635	0.91	0.91	MetaMetrics
	5 synthesizing memos from multiple sources	1239	0.88	0.79	Council for Aid to Education
Spoken	2000 spoken English tests adults, diverse items types	50	0.97	0.98	Balogh & et al. (2005)
	3000 spoken Arabic (diverse item types)	134	0.97	0.99	Bernstein et al. (2009)
	9 Oral Reading Fluency passages for 1^{st} – 5^{th} grade	248	0.98	0.99	Downey et al. (2011)

Table 1. *Indicators of scoring quality for six operational item sets. N is the number of test-takers per test or item used in the calculations. Machine-Human correlations are between one fully automatic score and a consensus human score (rating the same material). For comparison, the Human-Human column shows correlations between scores from two human graders for the same materials.*

For example, the third row of Table 1 shows score accuracy indicators for a set of five information-integration items. For each item, students were asked to write memos that synthesized information from multiple sources, including letters, memos, summaries of research reports, newspaper articles, maps, photographs, diagrams, tables, charts, and interview notes or transcripts. The resulting set of 1239 written responses was then scored by machine and by independent raters. The 0.88 correlation coefficient in the "Machine-Human" column indicates that machine scores are a reasonably close match to a good human score derived by combining the human scores to form a stable consensus score. One can intuit how close a 0.88 correlation is for these data by comparing it to the average correlation between pairs of scores given by pairs of individual skilled human raters, shown as the Human-Human correlation of 0.79. Thus, for these memos, automatic scores are closer to a stable consensus human score than one expert score is to another.

Performance Types to Score

Setting aside handwritten text and freehand graphic responses, which as yet cannot be scored automatically, constructed response item types amenable to automated scoring can be organized by presentation and response modality as represented in the following two matrices:

Scoring Focus	Presentation		Response	
	Spoken	Written	Spoken	Written
Declarative Knowledge	+	+	+	+
Language Skills	+	+	+	+

Scoring Focus	Presentation			Response		
	Graphic Info	Written Text	Equation and/or Graphic	Graphic Info	Written Text	Eq. & Graphic Editor
Mathematics	+	+	+	+	+	+

For measurement of both knowledge and language skills, Pearson has fielded automatic scoring within instructional products and in educational testing. Automatic scoring has already worked well in a number of performance task types including, among others:

1. Passage summary (written or spoken presentation and written or spoken response),
2. Written form (essay, email, response to a scenario, quality of group problem solving),
3. Spoken form (read aloud, retell passages, elicited imitation, reconstruct sentences),
4. Short-answer questions (written or spoken presentation and response), e.g. reading comprehension, science concepts, multisource information synthesis,
5. Oral reading fluency (rate, accuracy, and expression).

Examples of these operational item types alone and in combination are briefly described in the following section.

Item Examples

Essay Writing

In the WriteToLearn® product, an essay prompt (e.g. writing a letter to a local pet store) is presented to the student, who writes a response in the textbox below the prompt. The panel on the right shows the feedback results given by the system. The system gives an overall score as well as scores on traits of writing. The student's previous scores are shown by the triangles above the blue score bars. Criterion scores are shown by the green bars. Analysis of spelling, grammar, and possible repetition of information is available by clicking on the links provided. Clicking on individual traits, such as content or organization provides more detailed explanations of how to improve particular scored aspects of writing.

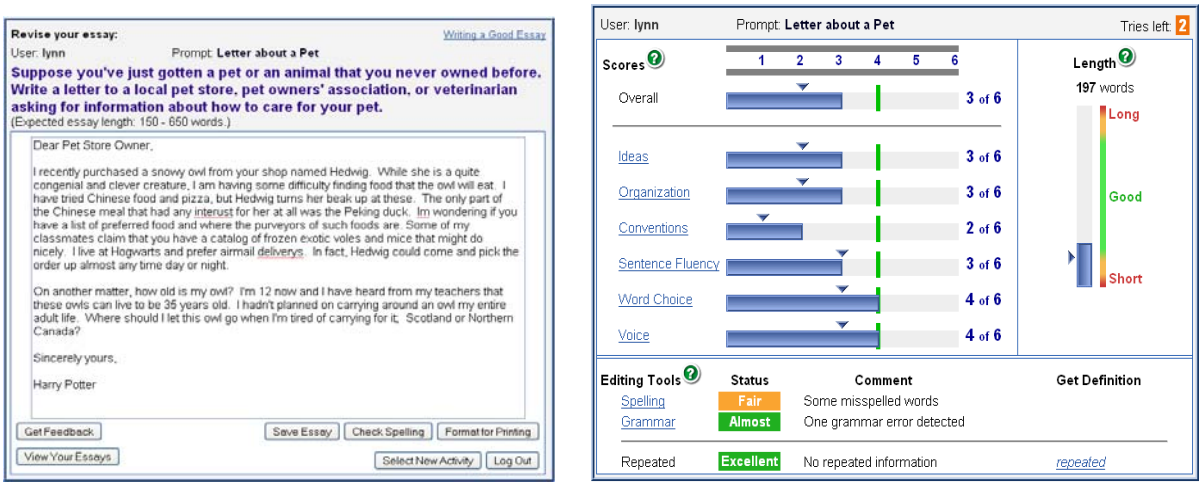


Figure 1: *Two screen shots from the WriteToLearn product. The left image shows a prompt and a typed-in response; the right image shows the student feedback scoreboard.*

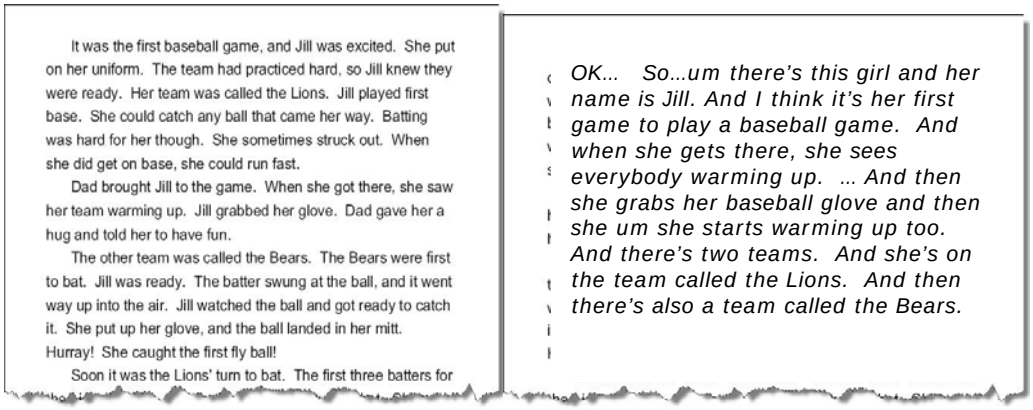
Curriculum Content	Presentation		Response	
	Spoken	Written	Spoken	Written
Declarative Knowledge		+		+
Language Skills				+

Oral Reading Fluency and Summarization

In several Pearson instructional products and assessments, a text passage is presented to students who read it aloud, then summarize the passage (narrative or expository) in their own words. When the response is spoken, depending on the application, the scoring system returns any or all of the following score types:

1. Oral reading rate, accuracy and/or expressiveness,
2. Decoding and word recognition skill,
3. Reading comprehension,
4. If English is the second language, vocabulary and/or proficiency level,
5. Fluency in spontaneous speaking.

In Pearson's Oral Reading Fluency system, students read a grade-appropriate passage for a minute. Pearson's KT system then analyzes the speech for rate, accuracy and fluency. The automatic scores correlate with human scores at 0.98, while the correlation between pairs of human raters is 0.99 (based on 245 delivered fluency tests delivered to 4th grade students). Below[i] is an example of a 4th grade passage used with the Oral Reading Fluency system and a transcript of the student's spoken summary response.



It was the first baseball game, and Jill was excited. She put on her uniform. The team had practiced hard, so Jill knew they were ready. Her team was called the Lions. Jill played first base. She could catch any ball that came her way. Batting was hard for her though. She sometimes struck out. When she did get on base, she could run fast.

Dad brought Jill to the game. When she got there, she saw her team warming up. Jill grabbed her glove. Dad gave her a hug and told her to have fun.

The other team was called the Bears. The Bears were first to bat. Jill was ready. The batter swung at the ball, and it went way up into the air. Jill watched the ball and got ready to catch it. She put up her glove, and the ball landed in her mitt. Hurray! She caught the first fly ball!

Soon it was the Lions' turn to bat. The first three batters for

OK… So…um there's this girl and her name is Jill. And I think it's her first game to play a baseball game. And when she gets there, she sees everybody warming up. … And then she grabs her baseball glove and then she um she starts warming up too. And there's two teams. And she's on the team called the Lions. And then there's also a team called the Bears.

Figure 2. *Left panel shows a read-aloud passage marked to reflect an oral reading performance. Right panel is a transcription of the same student's spoken summary.*

Curriculum Content	Presentation		Response	
	Spoken	Written	Spoken	Written
Declarative Knowledge				
Language Skills		+	+	

Summarization for Assessing Reading Comprehension

In the WriteToLearn product, reading comprehension is assessed by the content of the student's summary shown in the right panel, generated after reading a chapter-length passage, shown in the left panel. Feedback is provided showing how well the student covered the content of each major section in the reading. Feedback also includes indications of student copying from the text, redundant and irrelevant information as well as whether the summary is too long or too short. Students can reread and revise their summary until reaching a predetermined passing threshold.

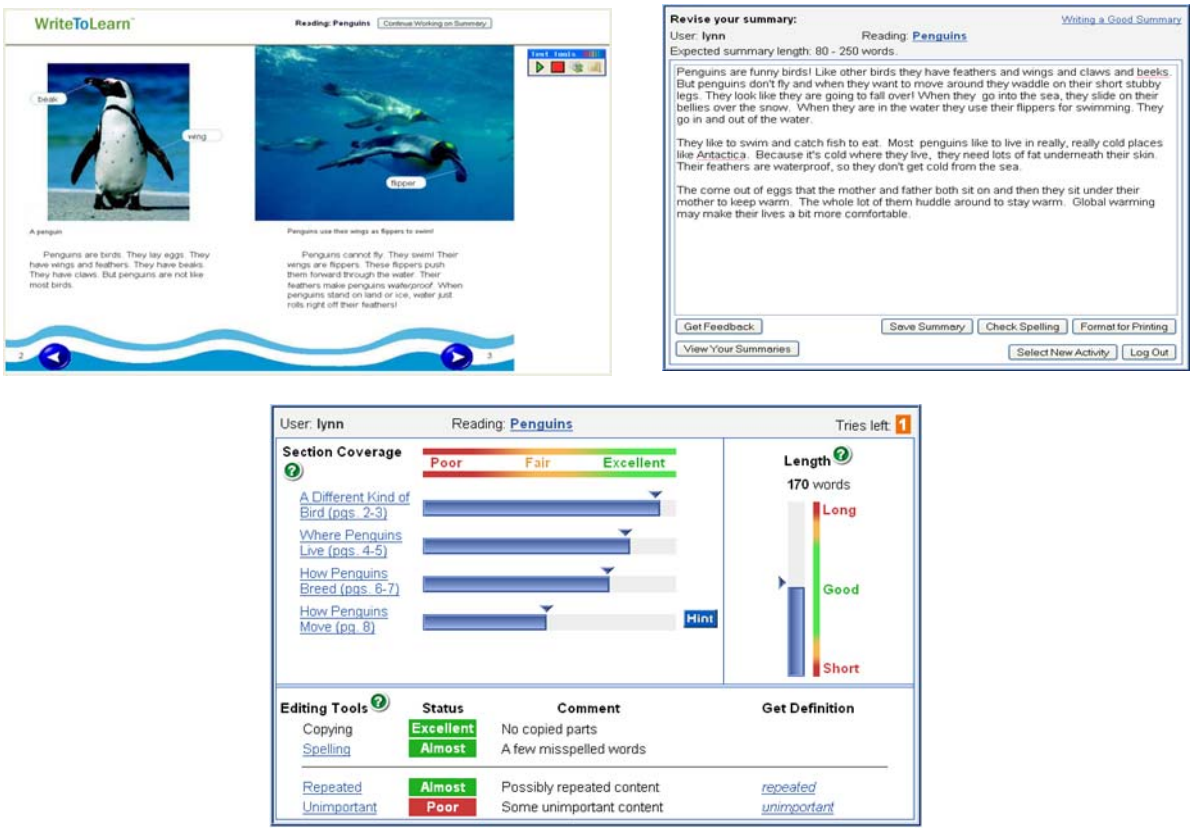


Figure 3: *Summarization task. The student reads the chapter about penguins shown in the upper left and writes a summary as shown in the upper right panel. The bottom panel shows the summarization score board. Content coverage on each of the four major sections of the reading is shown by the blue bars. The student's previous submission is shown by blue triangles. The length of the summary is in the "good" region—neither too short nor too long.*

Curriculum Content	Presentation		Response	
	Spoken	Written	Spoken	Written
Declarative Knowledge		+		+
Language Skills				+

Core Spoken Language Proficiency

Pearson offers Versant™ spoken language proficiency tests for English, Spanish, Dutch, Arabic, and soon Mandarin and French. In the Versant English Test either spoken or written items are presented. The student hears a sequence of three short phrases and responds by re-ordering the elements into a sensible sentence. The panel on the left shows the text of a sentence-build item. Combining base measures from five item types (read sentences, repeat sentences, answer short questions, build sentences from phrases, and re-tell spoken passages), the scoring system returns an Overall speaking measure, composed from four reported subscores: Sentence Mastery, Vocabulary, Fluency, and Pronunciation.

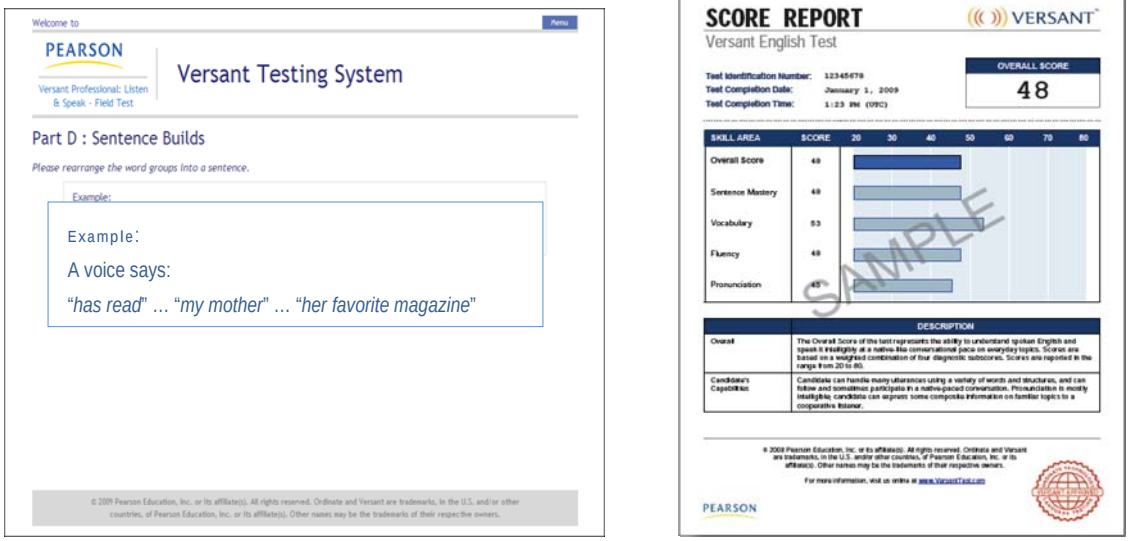


Figure 4. *The left panel is a screen showing an example sentence-build item. Following one example presentation by voice, which is printed on the form shown as well, the test taker speaks the response. All items are presented and responded to only by voice. The right panel is a sample score report for an L2 spoken English test.*

Curriculum Content	Presentation		Response	
	Spoken	Written	Spoken	Written
Declarative Knowledge				
Language Skills	+	+	+	

Short Answer Scoring for Science Concepts

Short answer science questions are scored automatically for the Maryland state assessments. Students read a text on a scientific question and write an answer after reading the relevant text. Shown here on the left is an example of a student's answer before reading a passage on magnetism. On the right is the same student's written response after reading the text. Short answers typically are between 5 and 50 words of text. The Pearson scoring engine returns partial- or whole-credit scores for science content. The technology can be broadly applied for any content domain requiring students to demonstrate their knowledge and explain their reasoning.

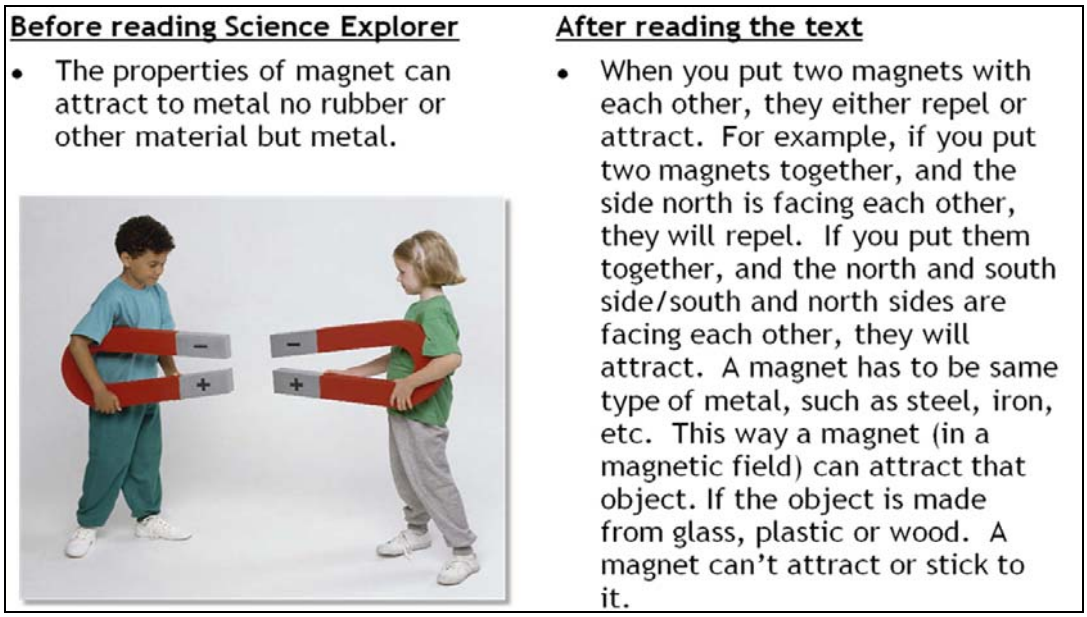


Figure 5. *An example of a science short answer. The answer on the left was given before reading the chapter on magnetism. The answer on the right was written after reading the chapter.*

Curriculum Content	Presentation		Response	
	Spoken	Written	Spoken	Written
Declarative Knowledge		+		+
Language Skills				

Four-Skills Language Assessment for Career and College Readiness

Two different four-skills tests assessing reading, writing, listening, and speaking implement a range of constructed response item types. One test is designed to qualify candidates for entry to academic programs, and the other is designed to qualify candidates for positions in business or in training programs for business.

Pearson's Knowledge Technologies group scores the *Pearson Test of English—Academic (PTE-A)*, which measures non-native speakers of English wishing to pursue higher education in English-speaking institutions. Depending on their length, written answers are scored by methods optimized for essays or for short answers. Spoken-response items are assessed for fluency and proficiency – again with somewhat different methods depending on the task. Test material was designed to reflect the kinds of language that are needed for university study. For example, in one type of item, a student hears a professor delivering a short lecture and is asked to give a spoken summary. Another item type asks for a one sentence written summary of an oral lecture snippet. The items measure language production, written and oral skills, pronunciation, and fluency among other language traits. For the response material elicited by this test, the average correlation between two human scorers was 0.87, while the correlation between a stable combined human score (a better estimate of the true score) and the machine score was 0.88.

Pearson's *Versant English Test–Professional* (Versant Pro) is another four skills test, which measures non-native speakers of English-speaking skills in the workplace. Versant Pro uses the same scoring technologies used in the PTE-A, but the test tasks reflect the kinds of language use needed in the commercial sector. For example, a test-taker reads a repair procedure as if to a customer, and composes emails addressing specific workplace issues. The items elicit language that provides evidence in support of reported score that cover:

Written Communication Skills	Spoken Communication Skills
OVERALL Grammar Vocabulary Organization Voice & Tone Reading Comprehension Suggestions to improve	OVERALL Sentence Mastery Vocabulary Fluency Pronunciation Listening Comprehension Suggestions to improve

Curriculum Content	Presentation		Response	
	Spoken	Written	Spoken	Written
Declarative Knowledge				
Language Skills	+	+	+	+

When scoring ***written*** responses on the Versant Pro test, the average correlation between two human scorers was 0.98, equal to the correlation between a combined human score and the machine score: 0.98. For ***spoken*** responses on the Versant Pro test, the average correlation between two human scorers was 0.99, while the correlation between combined human score and the machine score was 0.95. The demonstrated accuracy of the PTE-A and the Versant Pro tests shows that KT's automated scoring technology can be used with confidence in complex high stakes assessments. Table 2 on the following page displays the item types presented in the Versant-Pro test.

Task	Task Description	Examples
WRITTEN A: Typing	Type a given passage exactly as displayed in 60 seconds. Assesses typing speed and accuracy.	Passage: *For over 50 years, one car company has been making a classic sports car. The car is mostly handmade in western England…*(etc)
WRITTEN B: Sentence Completion	A sentence appears with a missing word. The test-taker has 25 seconds to fill the blank with a fitting word.	It's _____ tonight; bring your sweater. The company was _____ over fifty years ago. Hundreds of people _____ our grand opening.
WRITTEN C: Dictation	Listen to a sentence and type it exactly as heard. 25 seconds is allotted to type the sentence as accurately as possible.	Why didn't you ask him to leave a message? We hope that this information has been helpful. We don't ship to locations outside of Canada.
WRITTEN D: Passage Reconstruction	A short paragraph to be read appears on the screen for 30 seconds, after which, the paragraph disappears. In 90 seconds, type a paraphrase with the main points and as many details as possible.	Example Paragraph: *We remind all employees that any problems with any of the office equipment should be directed to the maintenance department. Please … get the equipment repaired as quickly as possible. Thank you.*
WRITTEN E: Email Writing	Read a description of a workplace situation and in nine minutes compose an appropriate email in reply. The reply must include three specific themes that are supplied to the test-taker as well as original supporting ideas for each theme.	*Workplace Situation*: You work for a non-profit organization that supports small businesses. You have been contacted by Mr. Roberts, who is interested in learning more about the organization. Include as themes: business network, newsletter, annual meeting.
SPEAKING A: Read Aloud	Read a passage aloud for 30 seconds. Candidates may not be able to finish reading the entire passage in 30 seconds, but this is not counted against them.	Many offices are becoming more and more diverse in the current … successful work environment is to appreciate each …also embrace diversity rather than deny differences between people.
SPEAKING B: Repeats	Repeat each sentence word-for-word. The sentences are presented in approximate order of increasing difficulty.	It took a lot longer than expected. Come to my office after class if you need help. People know how easy it is to get lost in thought.
SPEAKING C: Short Answer Questions	Listen to spoken questions and then answer each question with a single word or short phrase.	How many sides does a triangle have? How many months are in a year and a half? Either Larry or Susan had to go, but Susan couldn't. Who went?
SPEAKING D: Sentence Builds	Listen to a sequence of three short phrases and then rearrange them into a sentence.	my boss / to California / moved the prices range / to thirty dollars / from fifteen
SPEAKING E: Story Retelling	Listen to a story and then describe what happened in your own words.	Paul planned on taking the late flight out of the city. He wasn't sure whether it would be possible because it was snowing quite hard. In the end, the flight was cancelled because there was ice on the runway.
SPEAKING F: Response Selection	Listen to a sentence followed by three possible responses, then select the response that is most appropriate.	Our profit last year was higher than expected. A: Great, let's celebrate. B: That's too bad. C: We lost a lot last year.
SPEAKING G: Conversations	Listen to a conversation between two speakers and then give replies to comprehension questions.	Speaker1: *How was the business trip?* Speaker2: *There was a storm the whole time.* Speaker1: *That sounds terrible.* Q: *What happened during the business trip?*
SPEAKING H: Passage Comprehension	After listening to a spoken passage, answer three comprehension questions about the passage.	Q1: What problem did George have at the store? Q2: Why did George go to the shoe department? Q3: What did George decide to do after he saw the note on the door?

Table 2: *Item tasks used in Versant-Pro with examples.*

Expression Manipulation with an Equation Editor

Using an equation editor, a student can perform unconstrained editing to express a response as a single fraction or as a sum of two fractions. Each criterion is scored independently—whether the student has removed the radical from the denominator, and whether the student's response is equivalent and contributes to the cumulative score. The lower panel shows the assessment criteria used for this problem.

When problem responses can take a variety of forms as in the example, automated scoring is more reliable than human scoring.

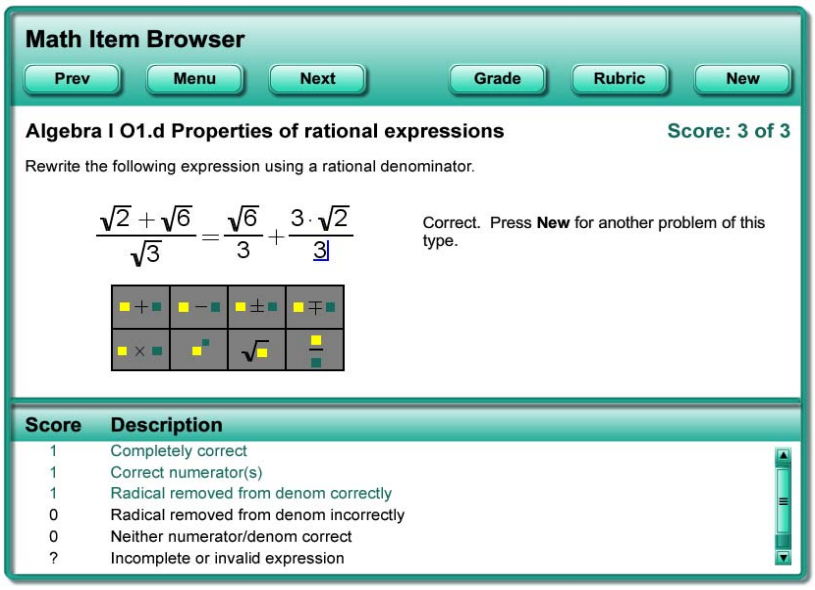


Figure 6. *Screen shot of a math item that uses MathQuery's editor interface and scoring against a rubric.*

Curriculum Content	Presentation			Response		
	Graphic Info	Written Text	Equation and/or Graphic	Graphic Info	Written Text	Eq. & Graphic Editor
Mathematics		+	+			+

Graphing Steps to Represent an Equation

Pearson's MathQuery™ supports math representations by using manipulatives that are designed specifically to capture the data that MathQuery needs. With graphing problems, for example, MathQuery can evaluate whether a curve is increasing or decreasing over specified intervals, and whether critical points like vertices, inflexion points, and axis intercepts are accurate to a specified tolerance (see Figure 7). In the example shown below, the system scores against a specified rubric returning three partial scores shown in the bottom panel.

Ongoing extensions to the MathQuery user interface will allow students to show and explain their work as they step through a computation, derivation, or proof, with evaluations at each student-identified stage of the process.

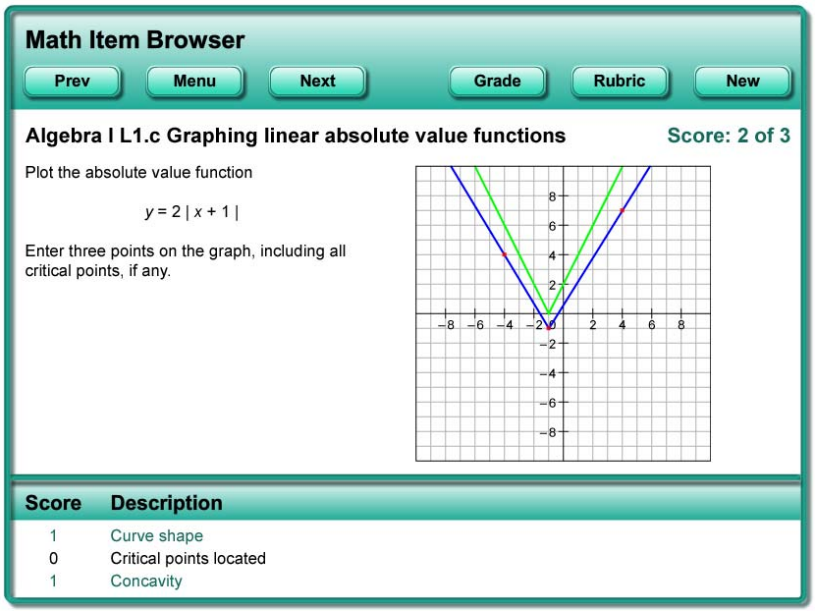


Figure 7. *An example of a graphing problem for which several aspects of the response can be independently scored.*

Curriculum Content	Presentation			Response		
	Graphic Info	Written Text	Equation and/or Graphic	Graphic Info	Written Text	Eq. & Graphic Editor
Mathematics		+	+	+		

Scoring Technologies

Scoring Writing

The following table displays four item types that are scored automatically in Pearson tests and/or instructional products. The item types differ in range of expected response, in data requirements during development, and in the traits measured.

Item Type	Response Length in Words	Typical Data Requirements for development	Measures Returned
Prompt-Specific Essays	100-500	200-250 double-scored student essays	Overall score, trait scores, grammar & mechanics feedback
Prompt Independent Essays (general models)	100-500	Approximately 1000 essays per grade	Overall score, select trait scores, grammar & mechanics feedback
Short Answers	~10-60	500 double-scored student answers	Total or partial-credit content score
Summaries	50-250	Readings to be summarized divided by major sections	Content coverage score for each section; checks copying, length, redundancy and irrelevance.

Table 3. *Characteristics of writing item types that are currently automatically scored*

Generally, essays are scored with reference to two kinds of criteria: form and content, both of which can be viewed in smaller or larger pieces. KT is the inventor of content-based scoring (see Landauer, Foltz, & Laham, 1998; Foltz, Gilliam & Kendall, 2000; Landauer, Laham, & Foltz, 2002) and has developed many educational applications of content analysis. KT also scores the form and style of written material.

Item-Specific Essay Scoring. KT scores essays by using statistical models that are based on expression of knowledge and on command of linguistic resources. An essay's content is measured by Latent Semantic Analysis (LSA), a statistical semantic model invented by KT's principals in the late 1980s. LSA is now in wide use around the world in many applications in many languages, including internet search, psychological diagnosis, signals intelligence, educational and occupational assessment, and in basic studies of collaborative communication and problem solving. Note that development of specifically calibrated essay scoring requires collection of scored response examples to construct a specific scoring model.

LSA derives semantic models of English from an analysis of large volumes of text equivalent to all the reading a student may have done through high school (about 12 million words). LSA builds a co-occurrence matrix of words and their usage in paragraphs, then reduces the matrix by Singular Value Decomposition (SVD), a technique similar to factor analysis. KT typically uses 300 independent numbers to represent the meaning of each word or paragraph in English. The accuracy of the LSA meaning representation is indicated by machine-human correlations in rating the similarity of meaning between pairs of paragraphs and the similarity of meaning between pairs of words. Evidence confirms that LSA rates similarity of meaning 90% as well as two human raters agree with each other about word and paragraph meanings (Landauer Foltz & Laham, 1998).

Other automatically computed variables are used to score the form and stylistic aspects of essays. Pearson's automatic system matches teachers' scores of traits such as voice, organization, word choice, and prescribed grammar. Measures based on raw length of essays, sentences or paragraphs are seldom included unless explicitly called for by a test design and documented for users.

Prompt Independent Scoring. When educators are mainly interested in gauging the stylistic and mechanical aspects of writing, KT has developed a generalized grading method (prompt-independent scoring) that is calibrated on thousands of essays across multiple topics and prompts. Prompt-independent scoring is an active research focus at KT, as it is still somewhat less reliable (self-consistent) than prompt specific scoring. A great advantage remains for formative instructional use however, because with prompt-independent scoring, teachers can easily author essay prompts that are tied to their own lesson plans and curriculum.

Short Answer Responses. Methods developed for scoring full essays need to be modified for short answer responses. In contrast with essays, the quality of short constructed responses is characterized more by word choice and the usage of specific terminology. To deal with these differences KT uses statistical classifiers and assessment-specific heuristics for treating ordering of events in a process or explanation to model each short answer. Compared to essay scoring the development of short-answer-response scoring requires more student data to reach the accuracy required for high stakes use. Based on research with the state of Maryland over five years, KT has found that about half to two thirds of the short-answer science items can be scored automatically with similar accuracies to human scorers. KT operates as a second scorer on those questions. For the remaining items, human scoring is used exclusively.

Instructional Summary Scoring. Summary writing lets students practice writing across content areas, providing instant content feedback by assessing their reading comprehension. The read, write, and revise cycle encourages students to reread, rethink and re-express those parts of the text that they have not yet fully understood. KT's summary feedback measures how well the student's writing has covered the content of each major section of a document by calculating the semantic similarity between the summary and each section of the text. Studies of its use in classrooms have shown that it produces improved reading comprehension and improved content writing when compared to students who did not receive automated feedback (Franzke et al. 2005).

Scoring Speaking

Pearson's spoken language technology is based on the recognition of spoken material and an analysis of that material with reference to words and other linguistic structures identified in the response. Computer scoring of spoken material starts with an augmented form of output from a speech recognition system that typically is optimized for the item that prompted the spoken response. This output usually includes a set of base measurements and tags that are aligned with a text string. The base measures and tags include:

1. A text string that represents the sequence of words spoken, along with any disfluencies;
2. A sequence of other linguistic units (phones, part of speech tags, phrases, and clauses) aligned both with the words in the text string and with the acoustic signal;
3. Figures of merit for each linguistic unit, including its duration and pronunciation quality.

From the base measures, computer scoring extracts many kinds of composite information:

1. Content, such as explicit or implicit propositions, and expository, procedural, or narrative sequencing;
2. Evidence of accepted word usage, collocation, and prescribed grammatical forms;
3. Prosodic information that conveys semantic focus and reflects comprehension.
4. Pronunciation and fluency for oral reading and second language assessment.

Example: Language Proficiency Scoring

Automated scoring for spoken language proficiency assessment is shown below.

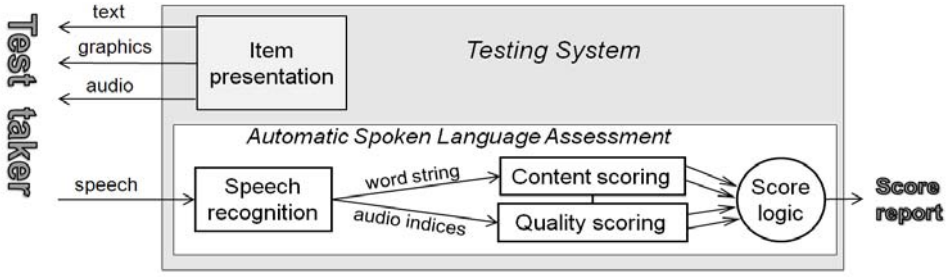


Figure 8. *Top-level view of spoken language assessment technology.*

A speech signal is picked up by a microphone and then processed by a speech recognition system that sends its output to scoring modules, which produce component scores that are combined into reported scores. When scoring second language proficiency, spoken response scoring focuses on the content of the speech, as well as the quality of the speaking. Content may include turn structure, pragmatic force, syntactic form and lexical choice. Scoring may also focus on qualities of the speech production itself such as its fluency and pronunciation, or it may combine content and quality aspects into a more general estimate of speaking ability. Current Pearson systems measure speaking ability in English, Spanish, and other languages by combining measures of linguistic and lexical structures with measures of fluency and pronunciation, returning scores with high consistency and accuracy. For example, the Versant Spanish Test, has a reliability of 0.98 and its scores correlate with certified interview tests with a coefficient of 0.92 (Bernstein, van Moere, & Cheng, 2010). Similarly, the spoken English responses from the Versant Professional test correlate with reference human scores at r = 0.95. More specific information on Pearson's scoring methods is described by Bernstein & Cheng (2008).

For summative assessment of spoken language ability, providing a variety of spoken tasks can return more accurate and reliable scores because the assessment is based on several minutes of speech and uses several independent sources of information. Four scores coming from the content and quality analyzers are schematized in Figure 9. These combine to produce an overall score that correlates closely with human summative judgments

Figure 9. *One set of Pearson spoken language proficiency scores*

Scoring Math

For mathematics, competency in math facts and operations is important, but real-world applications require a deeper understanding of mathematical principles that allows students to recognize what methods and operations are appropriate in different situations. Assessing these higher-order skills calls for performance tasks in which students are presented with a problem in context and have to determine which mathematical models are relevant, and further justify their reasoning.

Pearson's MathQuery (see Deland, 2007; Dooley, 2007; Hodgins et al., 2005) is a web-based environment that exercises and assesses critical thinking skills in math. These skills are best measured by multistep and real-world problems that can be solved more than one way and that can have multiple valid solutions that are not equivalent. MathQuery brings together technologies that display high-quality math notation and that allow student input of well-formed math responses. Criterion-based assessment algorithms automate important aspects of human scoring that go beyond numerical equivalence scoring. MathQuery generates and scores classes of algebraic and graphic problems using item schemas.

Figure 10 shows a sample floor-plan item in which students are asked to calculate the total area. MathQuery presents a floor plan, allows the student to enter an algebraic expression, and the scoring engine infers how the student calculated the area from the expression. Just evaluating the numeric value of the solution to see if it is correct, would shed no light on the problem solution process.

In this floor-plan example, MathQuery needs to determine which sub-expressions correspond to which sub-areas of the floor plan, which is done by pairing terms in the sum with particular sub-areas of the floor plan. In an instructional mode, the interface highlights areas in the diagram to help explain any errors found. MathQuery is able to perform this analysis because it captures and analyzes the expression structure provided by the student input. Note that in the floor-plan example, the student can solve the problem with either an additive or subtractive method (or a combination of both), which MathQuery can identify.

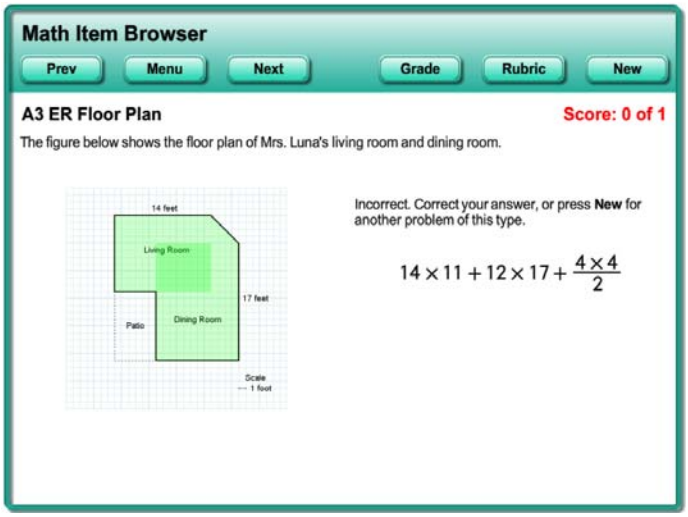


Figure 10. *A MathQuery presentation to solve for the area of a floor-plan problem.*

Note that this kind of math problem has multiple paths to the correct answer. In order to provide formative feedback and/or give partial credit, MathQuery analyzes the sequence of steps or the path to the solution. For mathematical expressions, MathQuery offers an equation editor that can be customized for different grade levels and content areas, so that pre-algebra students can easily express fractions, but are not overwhelmed by the functionality and symbols needed for calculus. In addition, the equation editor can correct input errors during response construction and if errors are not caught during input, MathQuery's assessment engine can accommodate input errors during grading by adapting the assessment criteria to the unexpected input.

Summary and Outlook

Automated scoring technologies already provide reliable and accurate methods for assessment of many kinds of constructed responses. Current methods are scalable and practical, allowing the development of novel formative and interim assessment tools as well as integration into benchmark and summative tests. Ongoing research and development will expand the range of item types that can be scored as well as improve the reliability of existing methods.

Pearson has created technology for scoring constructed responses and has graded many millions of such responses over the last 15 years. This experience and technology shows that many requirements for teaching and assessing listening, speaking, reading, and writing skills with authentic assessments can now be met with automatic systems. Pearson has a prototype of automated scoring of multi-step mathematical problems. In order to refine the student interface and provide validity evidence, Pearson's equation editor and math semantics analyzer still need to be field tested and validated.

The written and spoken items described above are all operational. Each day, Pearson's computers process and score hundreds of thousands of spoken and written responses from around the world.

Pearson's written technology is now used for operational scoring of:

1. College Board's ACCUPLACER test and the Pearson Test of English – Academic,
2. Practice essays for publishers and test preparation companies,
3. Essays and summaries for Pearson's WriteToLearn product,
4. Short answers for Maryland science tests.

Pearson's spoken response scoring technology supports reading fluency assessment and spoken language proficiency measurement in four languages so far, including:

1. English for universities and companies in the U.S. and abroad,
2. Dutch for the Netherlands Ministry of Justice,
3. Spanish for the U.S. Border Patrol Academy,
4. Arabic for the Army's Defense Language Institute.

Pearson's Knowledge Technology group looks forward to the opportunities and challenges to be met in the transformation of instruction and assessment in American schools.

References

Balogh, J., Bernstein, J., Cheng, J., Townshend, B., and Suzuki, M. (2005). *Ordinate Scoring of FAN in NAAL Phase III: Accuracy Analysis.* Menlo Park, California: Ordinate Corporation. 61 pages. Project Final Report to NCES. Shorter version: 'Automatic evaluation of reading accuracy: Assessing machine scores". Paper presented at the *Speech and Language Technology in Education Workshop*, (2007) Farmington, PA.

Bernstein, J., Suzuki, M., Cheng, J. and Pado, U. (2009). Evaluating diglossic aspects of an automated test of spoken modern standard Arabic. *ISCA International Workshop on Speech and Language Technology in Education (SLaTE 2009).*

Bernstein, J. and Cheng, J. (2008). "Logic, Operation, and Validation of the PhonePass SET-10 Spoken English Test," a chapter in V.M. Holland & F.P. Fisher, (Eds.) *Speech Technologies for Language Learning*. Lisse, NL: Swets & Zeitlinger.

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

Deland, D. (2007). An RIA Approach to Web Mathematics, presented at AMS/MAA Joint Mathematics Meeting, San Francisco, January 2010.

Dooley, S. S. (2007). MathEX: A Direct-Manipulation Structural Editor for Compound XML Documents. In *Proceedings of the Mathematical User-Interfaces Workshop 2007*, Schloss Hagenberg, Linz, Austria.

Downey, R., Bernstein, J. Cheng, J., and Rubin, D. (in preparation). Automating oral reading fluency for K-12 assessment.

Foltz, P. W., Gilliam, S. and Kendall, S. (2000). Supporting content-based feedback in online writing evaluation with LSA. *Interactive Learning Environments, 8*(2), pp. 111-129.

Franzke, M., Kintsch, E., Caccamise, D., Johnson, N., and Dooley, S. (2005), Computer Support for Comprehension and Writing. *Educational Computing Research*. 33(1), 53-80.

Hodgins, W., Dooley, S., Duval, E., and Lewis, S. (2005). Learning Technology Standards Committee (LTSC), *IEEE Standard for Learning Technology–Extensible Markup Language (XML) Schema Definition Language Binding for Learning Object Metadata* Standard 1484.12.3-2005. New York: IEEE Computer Society.

Landauer, T.K., Foltz, P. W., and Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes, 25,* 259-284.

Landauer, T. K., Laham, D., and Foltz, P. W. (2002). Automated scoring and annotation of essays with the Intelligent Essay Assessor. In M.D. Shermis and J. C. Burstein (Eds.) *Automated essay scoring: A cross-disciplinary perspective.* Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Topol, B. Olson, J. and Roeber, E. (2011). *The cost of new higher quality assessments: A comprehensive analysis of the potential costs for future state assessments*. Stanford Center for Opportunity Policy in Education.

Williamson, D. M., Bennett, R., Lazer, S., Bernstein, J., Foltz, P. W., Landauer, T. K., Rubin, D., Way, D., and Sweeney, K. (2010, June). *Automated Scoring for the Assessment of Common Core Standards.*