

PTE Academic automated scoring

March 2009

Introduction

Universities, higher education institutions, government departments and other organizations are increasingly faced with the need for an English language proficiency test that will accurately measure the communication skills of international students in an academic environment. In response to this need, Pearson Test of English Academic (PTE Academic) is being developed. The new test from Pearson will reliably measure the reading, writing, listening and speaking abilities of test takers who are non-native speakers of English and who want to study at institutions where English is the principal language of instruction.

Launching globally in 2009, PTE Academic will be offered in collaboration with the Graduate Management Admission Council® (GMAC®). GMAC is well known worldwide as the owner of the Graduate Management Admission Test® (GMAT®). In addition, PTE Academic will be delivered in a phased approach through Pearson VUE's test centers in a variety of countries. Pearson VUE is the global leader in electronic testing for regulatory and certification boards, providing a full suite of services from test development to test delivery to data management.

As the worldwide leader in publishing and assessment for education, Pearson is using several of its proprietary, patented technologies to automatically score test takers' PTE Academic performance. Academic institutions, corporations and government agencies around the world have selected Pearson's automated scoring technologies to measure the abilities of students, staff or applicants. Pearson customers using automated spoken and written assessments include eight of the 2008 Fortune Top 20 companies; 11 of the 2008 Top 15 Indian BPO companies; the U.S., German and Dutch governments; world sports organizations, such as the FIFA (organizers of the World Cup) and the Asian Games; major airlines and aviation schools; and leading universities and language schools.

An extensive field test program was conducted to test PTE Academic's test items and evaluate their effectiveness as well as to obtain the data necessary to train the automated scoring engines to evaluate PTE Academic items. Over the past 18 months, test data were collected from more than 10,000 test takers from 38 cities in 21 countries who participated in PTE Academic's field test. These test takers came from 158 different countries and spoke 126 different native languages, including (but not limited to) Cantonese, French, Gujarati, Hebrew, Hindi, Indonesian, Japanese, Korean, Mandarin, Marathi, Polish, Spanish, Urdu, Vietnamese, Tamil, Telugu, Thai and Turkish. The data from the field test were used to train the automated scoring engines for both the written and spoken PTE Academic items.

This paper provides a description of the automated scoring engines used in scoring PTE Academic and information about how and why they are accurate measures of written and spoken test taker performance.

Why automated scoring?

Research supports that, in many ways, automated scoring gives more analytical, objective results than humans do. Unlike human judgment, which is prone to be influenced by a variety of factors, an automated scoring system is impartial. This means that the system is not “distracted” by language-irrelevant factors such as test takers’ appearance, personality or body language (as can happen in spoken interview tests). Such impartiality means that test takers can be confident that they are being judged solely on their language performance, and stakeholders can be confident that a test taker’s scores are “generalizable” – that they would have earned the same score if the test had been administered in Beijing, Brussels or Bermuda.

Also, automated scoring allows individual features of a language sample (spoken or written) to be analyzed independently, so that weakness in one area of language does not affect the scoring of other areas. Human raters often exhibit “transfer of judgment” from one area of language to another. For example, test takers who speak smoothly may be marked as proficient even though their grammar is very poor. Automated scoring, on the other hand, assesses the different language skills objectively.

When developing its automated scoring technologies, Pearson conducts “validation studies” to make sure that the machine’s scores are comparable to scores given by skilled human raters. In a validation study, a new set of test takers’ responses (never seen by the machine) is scored by both human raters and by the automated scoring system. During Pearson’s validation studies, when the human scores are compared with the machine scores, they are found to be similar. In fact, the difference between the human score and the machine score is so small that it is usually less than the difference between one human score and another human score. This is true for both written and spoken assessments.

Research shows that the automated scoring technology underlying PTE Academic produces scores comparable to those obtained from careful human experts who are trained to consider only relevant language skills. This means that 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.

Scoring written English skills

The written portion of PTE Academic will be scored using the Intelligent Essay Assessor™ (IEA), an automated scoring tool that is powered by Pearson’s state-of-the-art Knowledge Analysis Technologies™ (KAT™) engine. Based on more than 20 years of research and development, the KAT engine automatically evaluates the meaning of text by examining whole passages. The KAT engine evaluates writing as accurately as skilled human raters using a proprietary application of the mathematical approach known as Latent Semantic Analysis (LSA). Using LSA, an approach that generates semantic similarity of words and passages by analyzing large bodies of relevant text, the KAT engine “understands” the meaning of text much the same as a human.

IEA can be tuned to understand and evaluate text in any subject area, and includes built-in detectors for off-topic responses or other situations that may need to be referred to human readers. Research conducted by independent

researchers as well as Pearson supports IEA's reliability for assessing knowledge and knowledge-based reasoning. IEA was developed more than a decade ago and has been used to evaluate millions of essays, from scoring student writing at the elementary, secondary and university levels to assessing military leadership skills.

Intelligent Essay Assessor and PTE Academic

IEA automatically evaluates a test taker's writing skills and knowledge and can be trained to score any writing traits that humans can reliably score. It assesses the total content of a test taker's response, using as a guide responses that were previously scored by expert human readers.

When taking PTE Academic, test takers will be asked to write 200- to 300-word essays and 50- to 70-word summaries. When a response is submitted for scoring, the system will evaluate the meaning of the response, as well as mechanical aspects of the writing. The system compares the response with the large set of training responses, computes similarities, and assigns a score based on content, in part by placing the response in a category with the most similar training responses. Scoring the mechanical aspects of the writing occurs in much the same way. The system assesses each trait (grammar, structure and coherence, etc.) in the test taker response, compares it with the large set of training responses, and then ranks the response according to that trait.

For the training of IEA, more than 50,000 written responses (essays and summaries) were collected in the field test. These written responses were scored on a number of traits including content, formal requirements, grammar, vocabulary, general linguistic range, spelling, development, structure and coherence. All test takers' responses in the field test were first scored by two human raters, and then by a third human rater when the first two did not agree. The scores from these human raters served as input for training IEA.

Because test takers' written responses were assigned randomly to raters drawn from a pool of more than 200 from Australia, the United Kingdom and the United States, the machine is trained on a rich set of international human judgments. The result is a person-independent rating. Based on the scores for all the traits mentioned above, an overall measure of writing performance can be formed by summing the trait scores for each test taker across all of the written items. This measure can be formed for the human raters and for the machine-generated scores. The correlation of these overall scores on this measure between pairs of human raters was 0.87. The correlation between the human score and the machine-generated score was 0.88. The reliability of the measure of writing in PTE Academic is 0.89.

Scoring spoken English skills

The spoken portion of PTE Academic will be automatically scored using Pearson's Ordinate technology. Ordinate technology is the result of years of research in speech recognition, statistical modeling, linguistics and testing theory. The technology uses a proprietary speech processing system that is specifically designed to analyze and automatically score speech from native and non-native speakers of English. In addition to recognizing words, the system locates and evaluates relevant segments, syllables and phrases in speech and then uses statistical modeling technologies to assess the spoken performance.

To understand the way that the Ordinate technology is “taught” to score spoken language, think about a person being trained by an expert rater to score speech samples during interviews. First, the expert rater gives the trainee a list of things to listen for in the test taker’s speech during the interview. Then the trainee observes the expert testing numerous test takers, and, after each interview, the expert shares with the trainee the score he or she gave the test taker and the characteristics of the performance that led to that score. Over several dozen interviews, the trainee’s scores begin to look very similar to the expert rater’s scores. Ultimately, one could predict the score the trainee would give a particular test taker based on the score that the expert gave.

This, in effect, is how the machine scoring is trained, only instead of one expert “teaching” the trainee, there are many expert scorers feeding scores into the system for each response, and instead of a few dozen test takers, the system is trained on thousands of responses from hundreds of test takers. Further, the machine does not need to be told what features of the speech are important; the relevant features and their relative contributions are statistically extracted from the massive set of data when the system is optimized to predict human scores.

While no human listener is likely to be accustomed to more than 100 different foreign accents, the speech processor for PTE Academic has been trained on more than 126 different accents and can deal with all of these accents equally. If the speaker has a very heavy accent and would be assigned a low score by typical human raters, then this test taker will receive a low pronunciation score from the machine. Importantly, the poor pronunciation would not influence the test taker’s grammar or vocabulary scores.

Ordinate technology powers the Versant™ line of language assessments, which are used by organizations such as the U.S. Department of Homeland Security, schools of aviation around the world, the Immigration and Naturalization Service in the Netherlands, and the U.S. Department of Education. Independent studies have demonstrated that Ordinate’s automated scoring system can be more objective and more reliable than many of today’s best human-rated tests, including one-on-one oral proficiency interviews.

Ordinate technology and PTE Academic

The Ordinate scoring system collects hundreds of pieces of information from the test takers’ spoken responses, such as their pace, timing and rhythm, as well as the power of their voice, emphasis, intonation and accuracy of pronunciation. It also recognizes the words that the speakers select (even if they are mispronounced) and evaluates the content, relevance and coherence of the response. Because the system is sensitive to many hundreds of linguistic and acoustic features in each response, it is able to provide a very precise estimate of how a skilled human rater would score each component of the response if paying specific attention to the component in question.

PTE Academic field testing provided data to create the automated scoring models for the spoken part of the test, just as it did for the written part. Nearly 400,000 spoken responses from the more than 10,000 test takers were collected. These included test takers’ spoken performances when describing figures or graphs, and re-telling lectures or presentations. Test takers’ responses were recorded and sent to human raters to be scored. Human raters scored test takers’ responses on a number of traits. The traits were content, vocabulary, language use, pronunciation, fluency and intonations. Aspects of the test takers’ responses objectively observable by the

advanced speech processing system, such as rate of speech, rhythm and word choice, were then compared with the raters' scores. Scoring models were then built, which are used to predict how trained human raters would score any "new" incoming response. The correlation between the human scores and the machine scores for an overall measure of speaking was 0.96. The reliability of the measure of speaking in PTE Academic.

When taking PTE Academic, test takers will be required to respond verbally to various kinds of questions. Their spoken responses will be captured as audio files and analyzed by the patented Ordinate scoring system. Some test questions will require short spoken responses. In these cases, the Ordinate scoring system measures the test taker's accuracy of word identification, pronunciation, fluency and grammatical facility. Other questions will be more complex, with test takers providing longer, more elaborate responses requiring many sentences or paragraph-level utterances. In addition to the traits listed above, the automated scoring system provides content and vocabulary scores on these responses.

Conclusion

By combining the power of a comprehensive field test, in-depth research and Pearson's proven, proprietary automated scoring technologies, PTE Academic will fill a critical gap by providing a state-of-the-art test that accurately measures the English language Speaking, Listening, Reading and Writing abilities of non-native speakers. Colleges, universities, government organizations and other institutions interested in becoming a PTE Academic-recognizing institution can visit www.pearsonpte.com or send an e-mail to the appropriate email address below for more information.

North and South America	usreco@pearson.com
Canada	canreco@pearson.com
United Kingdom and Ireland	ukireco@pearson.com
Europe, Middle East, African and India	emaireco@pearson.com
Asia-Pacific, Australia and New Zealand	apacreco@pearson.com

Bibliography

About Knowledge Analysis Technologies (KAT) Engine, Latent Semantic Analysis (LSA), and Intelligent Essay Assessor (IEA)

- Calfee, R. (2000). To grade or not to grade. *IEEE Intelligent Systems* 15(5), 35-37.
<http://www.pearsonkt.com/papers/IEEEdedebate2000.pdf>
- Landauer, T.K., D. Laham, & P.W. Foltz. (2003). Automatic essay assessment. *Assessment in Education: Principles, Policy & Practice*, 10(3), 295-308.
- Landauer, T.K., D. Laham, & P.W. Foltz. (2000). The Intelligent Essay Assessor. *IEEE Intelligent Systems* 15(5), 27-31.
- Landauer, T.K., P.W. Foltz, P. & D. Laham. (1998). Introduction to Latent Semantic Analysis. *Discourse Processes*, 25, 259-284. <http://www.pearsonkt.com/papers/IntroLSA1998.pdf>
- Landauer, T.K., & S.T. Dumais. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 104, 211-240.
<http://www.pearsonkt.com/papers/plato/plato.annotate.html>
- Pearson (2008). Reliability and Validity of the KAT Engine.
<http://pearsonkt.com/researchVRSum.shtml>

About Ordinate technology and Versant tests

- Bernstein, J., J. De Jong, D. Pisoni, & B. Townshend. (2000). Two experiments on automatic scoring of spoken language proficiency. In P. Delcloque (Ed.), *Proceedings in InSTIL2000*, pp. 57-61. Dundee, Scotland: University of Abertay.
<http://pearsonpte.com/research/automatedscoring>
- Harcourt (2006). Predicting ICAO levels from Versant™ for English. Author.
<http://harcourtassessment.com/hai/images/dotcom/vaet/ICAOPredictionFromVersant.pdf>
- Kerkhoff, A., P. Poelmans, J. de Jong, & M. Lennig (2005). *Verantwoording Toets Gesproken Nederlands*. [Account of the Test of Spoken Dutch] Den Bosch: CINOP.
- Pearson (2008). *Versant English Test: Test Description and Validation Summary*
<http://pearsonpte.com/research/automatedscoring>
- Pearson (2008). *Versant Aviation English Test: Test Description and Validation Summary*. Author.
<http://pearsonpte.com/research/automatedscoring>
- Pearson (2008). *Versant Spanish Test: Test Description and Validation Summary*
<http://pearsonpte.com/research/automatedscoring>
- Pearson (2004). *Versant English Test: Can do Guide; Ordinate® SET-10®*. Author.
<http://pearsonpte.com/research/automatedscoring>

About Pearson

Pearson (NYSE:PSO), the global leader in education and education technology, reaches and engages today's digital natives with effective and personalized learning, as well as dedicated professional development for their teachers. This commitment is demonstrated in the company's investment in innovative print and digital education materials for pre-K through professional learning, student information systems and learning management systems, teacher development, career certification programs, and testing and assessment products that set the standard for the industry. The company's respected brands include Scott Foresman, Prentice Hall, Addison Wesley, Benjamin Cummings, the Stanford Achievement Test, the Wechsler family of assessments, SuccessNet, MyLabs, PowerSchool, SuccessMaker and many others. Pearson's comprehensive offerings help inform targeted instruction and intervention so that success is within reach of every student at every level of education. Pearson's commitment to education for all is supported by the global charitable giving initiatives of the Pearson Foundation. Pearson's other primary businesses include the Financial Times Group and the Penguin Group.

For more information, go to www.pearson.com