Pearson
PTE Academic

Pearson Test of English Academic: Automated Scoring

January 2019

Contents

Contents

Introduction 2

Why automated scoring? 2

Scoring written English skills 3

Intelligent Essay Assessor and PTE Academic 4

Scoring spoken English skills 4

Ordinate technology and PTE Academic 6

Conclusion 7

References 7

- About Knowledge Analysis Technologies (KAT) Engine, Latent Semantic Analysis (LSA), and Intelligent Essay Assessor (IEA) 7
- About Speech Proficiency 7
- About Ordinate technology and Versant tests 8

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) has been developed. The test reliably measures 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.

Launched globally in 2009, PTE Academic is 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 is delivered globally through Pearson's test centers. 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' performance on PTE Academic. 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 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 tests. These test takers came from over 126 different countries and spoke more than 90 different 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.

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 a test taker’s 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 taker 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 is 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 elementary, secondary and university level, 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 responses that were previously scored by expert human readers as a guide.

When taking PTE Academic, test takers are asked to write 200–300 word essays and 50–70 word summaries. When a response is submitted for scoring, the system evaluates 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's 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 is 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 spoken performance.

In face-to-face spoken assessment situations, it is often the case that the candidate's language ability in sociolinguistic competence, pragmatic competence, strategic competence, conversation management, and/or turn-taking management tends to receive more attention than the intended assessment focus. While there is no denying that competency in these areas is important, much of the research indicates that the main underlying factor that unlocks the usefulness of these competencies is the candidate's processing competence (e.g., Skehan, 1998;) or the efficiency of processing (e.g. Van Moere, 2012). From the listening perspective, unless the candidate has the ability to efficiently recognize sounds, access and retrieve lexical and syntactic information, extract the interlocutor's intended meaning in real time, the candidate will likely face listening comprehension challenges in any spoken communications including face-to-face face interactions. Subsequently from the speaking perspective, the candidate then has to come up with the message or conversation point to communicate, access his/her mental lexicon and make lexical and syntactic decisions, and then articulate it in spoken sentences, as theorized by Levelt (1989). If the candidate cannot perform most, if not all, of these "psycholinguistic" processes in real time, it can be said that the candidate cannot deploy all other "sociolinguistic" or "pragmatic" competence. This academic research on the key factors in spoken language proficiency enabled Pearson to develop, test and verify automated speech recognition software that assesses the construct relevant aspects of speech with very high accuracy and without bias.

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 rater 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 is trained to score 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. Furthermore, 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 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 included content, vocabulary, language use, pronunciation, fluency and intonation. Aspects of the test takers' responses, which were 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 thus proving the reliability of the measure of speaking in PTE Academic.

When taking PTE Academic, test takers are required to respond verbally to various kinds of questions. Their spoken responses are captured as audio files and analyzed by the patented Ordinate scoring system. Some test questions require short spoken responses. In these cases, the Ordinate scoring system measures the accuracy of the test taker's word identification, pronunciation, fluency and grammatical facility. Other questions are 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 fits 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.

For further information about PTE Academic visit www.pearsonpte.com or email PLTsupport@pearson.com. For more information about automated scoring, we have a number of videos on YouTube.

References

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

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.

About Speech Proficiency

Cutler, A. (2003). Lexical access. In L. Nadel (Ed.), *Encyclopedia of Cognitive Science. Vol. 2, Epilepsy – Mental imagery, philosophical issues about*. London: Nature Publishing Group, 858-864.

Jescheniak, J.D., Hahne, A. & Schriefers, H.J. (2003). Information flow in the mental lexicon during speech planning: Evidence from event-related brain potentials. *Cognitive Brain Research*, 15(3), 261-276.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.

Levelt, W.J.M. (2001). Spoken word production: A theory of lexical access. *PNAS*, 98(23), 13464- 13471.

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford: Oxford University Press.

Vinther, T. (2002). Elicited imitation: A brief overview. *International Journal of Applied Linguistics*, *12*(1), 54–73.

Van Moere, A. (2012). A psycholinguistic approach to oral language assessment. *Language Testing* 29(3), 325–344

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.

https://pearsonpte.com/wp-content/uploads/2018/05/InSTiL2000_paper_Two_Experiments_Automatic_Scoring_Spoken_Language_Proficiency.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.

https://zoek.officielebekendmakingen.nl/kst-29700-30-b3.pdf

Pearson (2004). Versant English Test: Can do Guide; Ordinate® SET–10®. Author.

https://pearsonpte.com/wp-content/uploads/2018/05/Guide_Versant_English_Scores_Can-Do_Guide_comparison_to_CEF-R-v2.pdf

Pearson (2008). Versant Aviation English Test: Test Description and Validation Summary. Author.

https://pearsonpte.com/wp-content/uploads/2018/05/Versant_Aviation_English_Test_Validation.pdf

Pearson (2008). Versant English Test: Test Description and Validation Summary.

https://pearsonpte.com/wp-content/uploads/2018/05/Versant_English_Test_Validation.pdf

Pearson (2008). Versant Spanish Test: Test Description and Validation Summary.

https://pearsonpte.com/wp-content/uploads/2018/05/Versant_Spanish_Test_Validation.pdf