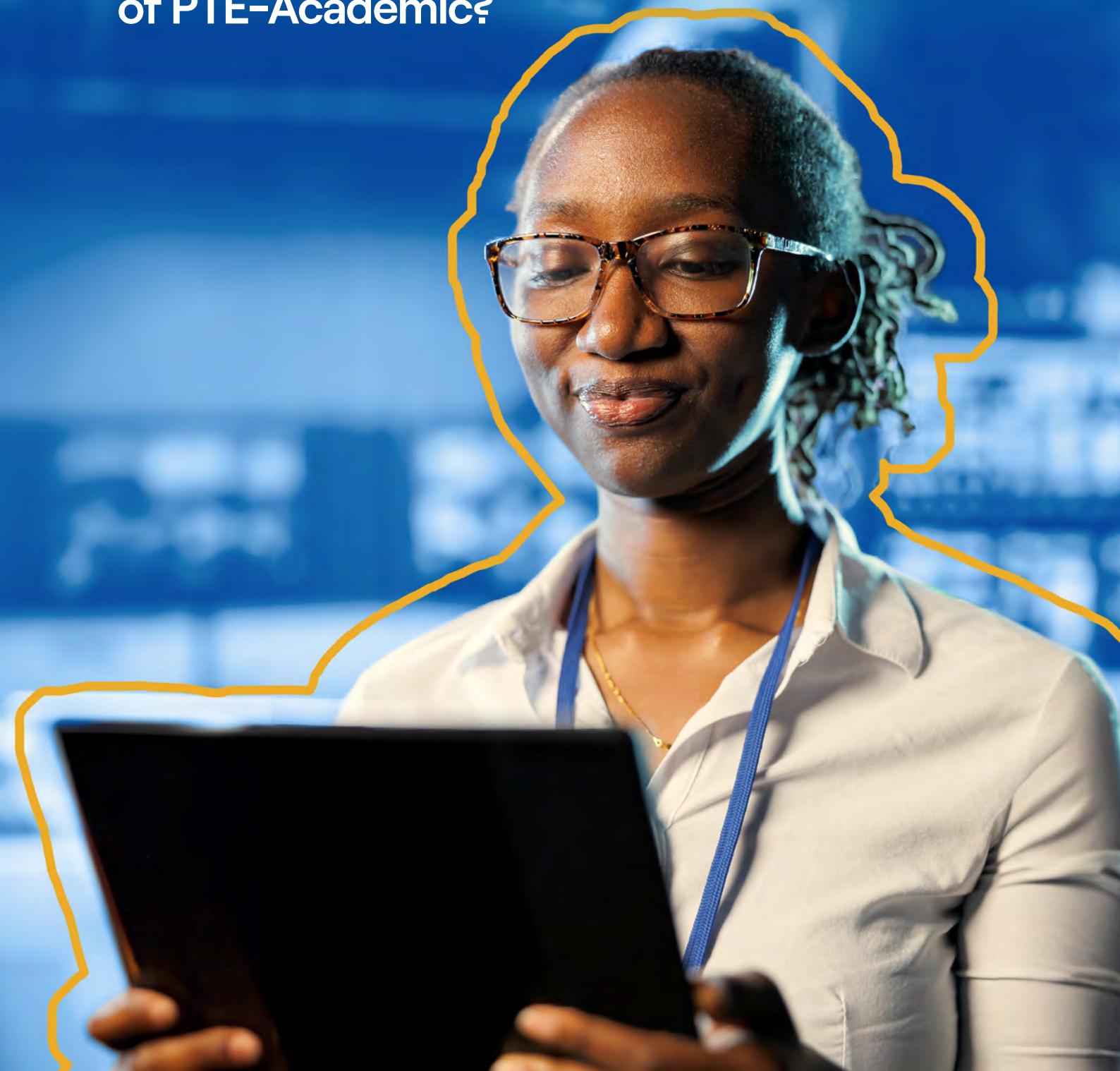
Pearson

Dr Bill Bonk

PTE AI Uncovered

How is AI used in the scoring of PTE-Academic?



Human examiners have traditionally been seen as the gold standard in language assessment. Who better to assess language performance than other expert users of that language? This may well be true, but the use of human examiners is not without issue.

– Dr Bill Bonk

PTE AI Uncovered: how is AI used in the scoring of PTE-Academic?

Artificial intelligence (AI) has been in the media quite a bit lately due to the development of new technologies that are starting to impact our daily lives, such as smart home devices, facial recognition software and even driverless vehicles. One of the greatest impacts in recent months has come from the launch of virtual assistants, such as ChatGPT, that can generate and analyze texts without human intervention, by using the massive amounts of data available in "large language models". Unsurprisingly, there has been a lot of media discussion around the implications of such technologies for education. How are students using this software? Are they using it as a "short cut" to learning? How can educators tell if an essay has been written by ChatGPT rather than by the student themselves?[1] Does this constitute cheating? Whilst most would now agree on the potential for this new iteration of AI to support and enhance education, I think it is fair to say that there is still a certain amount of suspicion surrounding these new developments and exactly what their impact will be on so many parts of our lives.

> As a result of all the media hype around OpenAI and ChatGPT, the term "Artificial Intelligence" is now seen by many as a synonym for these new types of technology – when in actual fact, the term itself covers a multitude of different technologies and approaches.

1 The impact of ChatGPT on education: The good and the bad

And when it comes to using AI to assess students, especially in high-stakes contexts, mystery and confusion abound! At Pearson Languages, we have been using AI to assess the proficiency of language learners for over 20 years – long before the release of ChatGPT and the recent popularity of AI tools simple enough for even children to experiment with. Therein lies the problem – and the confusion. As a result of all the media hype around OpenAI and ChatGPT, the term "Artificial Intelligence" is now seen by many as a synonym for these new types of technology – when in actual fact, the term itself covers a multitude of different technologies and approaches.[2]

In this piece I will dig into the much more controlled type of AI used in Pearson language assessments – a version of AI that draws extensively on human expertise in the field of language proficiency - and how its use addresses fundamental issues of fairness, accuracy and consistency. But before we get into this, let's unpack some of the particularities of language testing.

Assessments in language testing

As students in school, we have all experienced different types of assessment, ranging from multiple-choice quizzes to longer assignments and essays. These different types of assessment all have their place in language learning, but their goals are different and each method brings its own set of benefits and drawbacks. This spectrum of tasks more or less range from more constrained and less authentic-looking to more open-ended and more authentic-looking.

> "In many ways, performance assessment is a preferable type of assessment for measuring language proficiency since it measures the type of activities that you would expect a language learner to perform in the real world, such as writing an essay or talking about a particular subject."

Multiple-choice type assessments can reveal a lot about a student's knowledge of the language and are popular because they are easy to score. There is only one correct answer, which all experts would agree on, and the answer is either right or wrong. On the downside, tests like this can only indirectly assess the kinds of language skills that are used in the real world. Some subjects and areas of knowledge lend themselves well to this type of selected response testing (eg. math or grammar), whilst others, such as the ability to speak or write in a foreign language, do not.

2 What is artificial intelligence (AI). Everything you need to know.

In order to assess a test taker's ability to use a language rather than demonstrate knowledge of its rules, we need to look to the other end of the language assessment spectrum and the realm of performance assessment. This type of assessment is much more open-ended than the multiple-choice type, but typically takes much longer and is usually scored by expert human examiners, based on a set of scoring criteria (known as scoring rubrics). In many ways, performance assessment is a preferable type of assessment for measuring language proficiency since it measures the type of activities that you would expect a language learner to perform in the real world, such as writing an essay or talking about a particular subject. The downside of this type of assessment is the fact that individual human examiners are typically required to assess how well a particular test taker has performed the task, and that brings with it a different set of challenges.

Human examiners have traditionally been seen as the gold standard in language assessment. Who better to assess language performance than other expert users of that language? This may well be true, but the use of human examiners is not without issue.

> "Even with all the training, it is virtually impossible to guarantee 100% consistency across all examiners around the world."

In order to ensure, fair, reliable and consistent scoring, examiners need to be trained and regularly monitored. This is logistically complex and all the more challenging for international assessments for which large teams of raters around the world need to be standardized in assessing test takers from hundreds of different language backgrounds. By training every single human examiner to use the same set of scoring criteria, the aim is to ensure that the test taker in China is scored exactly the same as the test taker in Brazil, even though they have different examiners. You can imagine how difficult this standardization process is. Humans are notoriously inconsistent – with each other and with themselves (across different test takers at different times of the day). Even with all the training, it is virtually impossible to guarantee 100% consistency across all examiners around the world. Although a large pool of trained examiners might be available to score all tests, in practice your responses would only be rated by a very small number (usually two) and an unlucky draw could mean the difference between passing and not passing the test based on that draw. Often when two independent raters are used, if their scores agree then it is considered correct, even when both those raters may have been too lenient or too severe.

Furthermore, humans are prone to bias and may well be influenced in their scoring by a test taker's appearance, nationality or accent – either consciously or unconsciously.[3] This introduces questions of fairness into the assessment process, as well-trained and well-meaning as the raters themselves may be.

> "Giving tests with several different item types enhances the measurement of the skills in the assessment, boosts reliability, and makes it more difficult for anyone trying to improve their score through test-taking strategies such as the memorization of potential answers."

That said, most high-stakes assessments today attempt to incorporate a range of different assessment types:

- Items that tend more towards objective, discrete answer types (such as multiple-choice and gap-fill), because they tend to be shorter and do not require human judgments. This means that many items can be included in the same test, increasing reliability and minimizing the time and effort needed to grade the answers.
- Items that tend more towards qualitatively rich but subjective performance assessment, because they more closely mirror real-world uses. However, these types of item are more time-consuming to score because they require human raters who themselves need to be trained and standardized. As a result, there are usually only a few such item types in any given assessment which can result in misleading scores since the test taker has limited opportunity to demonstrate proficiency. Human raters may also be inconsistent.

In this way, assessments reap the benefits of having a range of different item types. Giving tests with several different item types enhances the measurement of the skills in the assessment, boosts reliability, and makes it more difficult for anyone trying to improve their score through test-taking strategies such as the memorization of potential answers.

3 Willis & Todorov. How many seconds to make a first impression. Journal of Psychological Science. July 2005.
Myford, C., & Wolfe, E. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. Journal of Applied Measurement, 4(4), 386–422.
Crooks, T., Kane, M., & Cohen, A. (1996). Threats to the Valid Use of Assessments. Assessment in Education: Principles, Policy & Practice, 3(3), 265–286.

> It is a form of responsible AI known as "machine learning" in which humans play a key role in training computer software to replicate their work.

The use of AI in PTE Academic

PTE Academic is a high-stakes assessment for academic study and professional migration. It sets out to measure a person's ability to use English in those contexts and for that reason it needs to include many performance-type assessment items. For more than 20 years, Pearson has been developing and harnessing new technologies to address the issues laid out above – namely the drawbacks of using human examiners. These new technologies fall under the umbrella term of Artificial Intelligence – but rather than relying on large language models (LLMs), as is the case for ChatGPT, they rely heavily on expert human ratings as part of the process to train the scoring engines.[4] This is not the same AI as used by OpenAI. It is a form of responsible AI known as "machine learning" in which humans play a key role in training computer software to replicate their work.

So how does "machine learning" work? In PTE A, we assess both spoken and written responses from test takers. Whilst the scoring of each is based on different traits or characteristics, the underlying methodology for training the scoring model is the same.

Step 1: Creation of an assessment framework

This phase involves writing a number of test items (questions) which address the topics that we want to cover. It also involves drawing up assessment criteria for the different traits that are to make up scores for the tasks. For example, on PTE A we measure traits such as content, grammar, vocabulary, coherence, fluency, and so on.[5] These criteria are established by applied linguistics experts and require a lot of discussion and analysis of language samples. In real time, it is not possible for the human brain to focus on individual traits and human examiners will tend to go for a more holistic approach to scoring. The human brain experiences something known as the "halo effect" which means that it is influenced by performance on other traits.[6] If a test taker is particularly strong or weak on the range of vocabulary they use, for example, this may influence the scoring of other traits and they could be marked up or down as a result. By contrast, a computer is capable of focusing on each individual trait with laser precision.

4 PTE scoring
5 PTE Academic Score Guide
6 Halo effects in grading: an experimental approach (Schmidt, F; Kaiser, A; Retelsdorf, J 2023)

> "In real time, it is not possible for the human brain to focus on individual traits and human examiners will tend to go for a more holistic approach to scoring."

Step 2: Collection of sample responses

Sample responses for all test items are collected from a representative sample of test takers from the target population (in the case of PTE A, test takers from a wide range of different countries, cultures and language backgrounds who are looking to study abroad or migrate for professional purposes). These responses are then scored by expert human examiners using the assessment criteria established in the previous step. These human-scored sample responses, at different levels of proficiency, form what is known as the "training set" – the set of responses used to train the scoring software during the machine learning phase. The larger and better the training set, the more rigorous the machine learning. PTE A's speech recognition software has been trained on over 400,000 spoken responses. Its Intelligent Essay Assessor used for scoring written texts has been trained on 50,000 essays.[7] By combining data from thousands of test taker responses and hundreds of human experts, we are able to ensure that the scoring of the live test is accurate and objective, rather than hinging on the opinion of a single human examiner.

> "By combining data from thousands of test taker responses and hundreds of human experts, we are able to ensure that the scoring of the live test is accurate and objective, rather than hinging on the opinion of a single human examiner."

7 Pearson test of English Academic: Automated Scoring (Pearson 2019)

Step 3: Training the computer scoring model

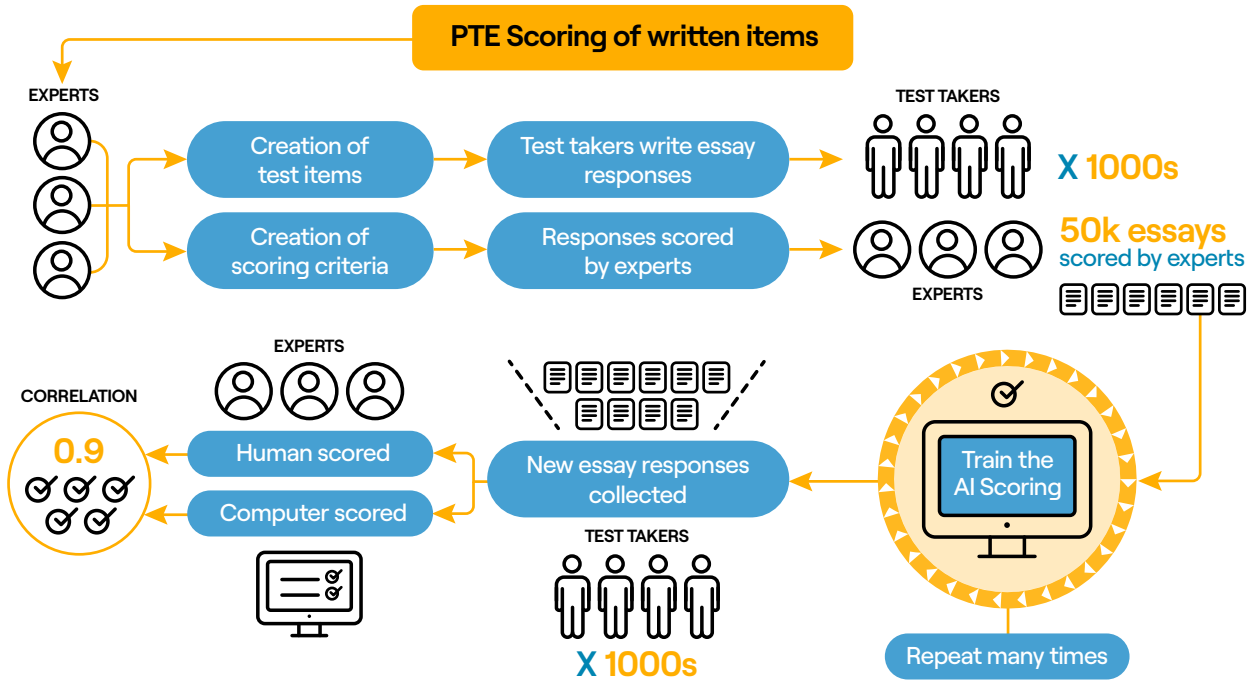
In the machine learning phase, the "training set" is used to fine-tune the computer scoring model for each test item. The computer does this by looking at the different traits that have been scored by the human examiners and trying to predict the human scores that were assigned for those learner responses by analyzing the many underlying "micro-features" which are automatically calculated behind the scenes. These underlying "micro-features" are the many aspects of language or speech that computer systems can measure quite accurately and automatically, such as the length of time in milliseconds between phrases in your spoken responses, or how well certain words "go together" in your written production.

> "After many iterations, the machine learning algorithm finally establishes the weighting that results in the closest match to the human scores."

The computer carries out the analysis of the "training set" over and over again, adjusting the weighting given to each of the different features, each time getting closer and closer to successfully matching the human scores assigned. After many iterations, the machine learning algorithm finally establishes the weighting that results in the closest match to the human scores.

Step 4: Testing the computer scoring model

The final phase is known as the "testing" phase. In this phase, the scoring model is given a large sample of fresh data (new learner responses to the same items) that it has never seen before. It uses the weights it calculated in the training phase to produce scores for the same set of traits. These same new learner responses are also scored by expert human examiners and the computer-scored responses are then compared with the human ratings. If the model is good at predicting the human scores, it is deemed a success and can be deployed in PTE A. If, for some reason, the model does not do a good job of predicting the human ratings, it is either re-trained until its results align with those of the human examiners, or the item is discarded. When looking at the correlation between human examiners and computer scoring, the analysis is expressed as a number between 0 and 1, where 0 means no correlation (very different results) and 1 means a strong correlation (exactly the same results). In many cases, the correlation between the computer scoring and the human scoring is higher than the correlation between two human scorers.



> "In many cases, the correlation between the computer scoring and the human scoring is higher than the correlation between two human scorers."

The role of second language experts in scoring

Experts in second language acquisition and use play a large role in many parts of the process described above. They develop the design of the test and consider the ways that different kinds of responses contribute to scores. They create the materials that guide item writers, the team recording audio scripts, and others whose contributions all end up being experienced by the test-taker. They create the "recipe" for how scores on different items and traits get combined into skill scores like Listening, Speaking, and so on. They train the human raters whose judgments are used to train the machine learning algorithms. They write the scoring rubrics that raters use to give scores in a consistent way. They guide the "standard-setting" process which takes scores on tests and links them to external standards of second language proficiency. They are the essential link between state-of-the-art knowledge and research on second language proficiency and the test scores that test-takers finally receive which are meaningful and decisive.

Human examiners v Machine learning v Large Language Models

The landscape of language assessment is evolving rapidly – as are the goals and needs of test takers and users. With greater implementation of new technologies, it is useful to summarize the pros and cons of the different approaches to measuring language proficiency.

Assessments using:	Pros	Cons
Human examiners	→ Trusted as being subject matter experts → Interaction with the test taker in spoken exams – a "human experience" → Flexible – can adapt the speaking test	→ Need for standardization training and ongoing monitoring of examiners → Potential bias of examiners → Inconsistency between examiners → Can be intimidating for test takers (power difference) → Time needed to provide test takers with results → Issues of test security → Impossible to focus on individual traits of spoken and written English – assessment is more holistic → 4 skills tested separately
Machine learning	→ Combines computer technology with huge numbers of expert human judgements → Accurate and consistent scoring → Can score individual traits in spoken and written language → Skills can be tested as integrated skills (reflecting real-world use of language) → Reduces test taker stress by removing the face-to-face oral exam → Quick results for the test taker → Highly secure → Rich data source for updates, item review, identifying fraudulent tests	→ Difficult to explain to non-experts how scores are given → No interaction in speaking items → Test takers can try to "trick" the computer → Limited number of spoken and written items (due to the requirements of machine training)
Large language models	→ Fewer (if any) constraints on the questions that can be asked in spoken and written items → Dynamic interaction in spoken items – closely mimicking interaction with another human	→ (As yet) unreliable and inconsistent– different software gives different results → Difficult/impossible to explain how a score was derived

> When it comes to high-stakes assessment – assessment that is of high value because it enables major life goals such as the ability to study abroad or migrate to a new country – institutions and governments around the world require reassurance that an assessment is a fair and accurate measurement of someone's language proficiency.

Integrating AI systems and human judgments

At Pearson, we are constantly monitoring test taker behaviors in order to ensure the integrity of the test and accuracy of test scores. Whilst most test takers are genuine, some engage in test preparation activities that are aimed at "gaming" the test - meaning they try to trick the computer scoring systems in order to achieve a higher score. One common approach is to memorize templated scripts for the more open spoken and written item types. As these memorized scripts become more and more sophisticated, it is increasingly difficult for computers or humans alone to identify them. Which is why Pearson is

tackling the issue through a combination of computer detection systems and expert human judgment. To find out more about the latest changes to the PTE A scoring systems, take a look at the paper written by my colleagues Sarah Hughes and Dr Rose Clesham on *Test Integrity: Ensuring truest by integrating human judgment and AI systems.*

Conclusion

The advances being made to technologies that draw on large language models are truly ground-breaking and will eventually disrupt the world of education and assessment. When it comes to high-stakes assessment – assessment that is of high value because it enables major life goals such as the ability to study abroad or migrate to a new country – institutions and governments around the world require reassurance that an assessment is a fair and accurate measurement of someone's language proficiency.

At Pearson Languages, our AI teams, researchers and assessment experts are constantly looking to the future and investigating ways in which to combine the best test taker experience with the most accurate measurement of language proficiency. Today we believe that we can achieve this by combining the strengths of both "machine learning" technologies and human expert input which together guarantee fair, consistent, and reliable results for every test taker around the world.

Dr Bill Bonk

About the author

Dr Bill Bonk is Director of Research Services of the PTE Assessmemt Research and Innovation group at Pearson. During his time at Pearson, Bill has developed a number of tests for both young learners and adults and his work has included the design and psychometric analysis of second language tests, and growth modeling of assessment data. Before working in test design, Bill taught English as a foreign language for over a decade in countries such as Italy, Ecuador, Japan, and Brazil. He has an M.A. in Second Language Studies and a PhD in Cognitive Science.