Pearson

Sarah R. Hughes and
Dr. Rose Clesham

# Test integrity

## Ensuring trust by integrating human judgment and AI systems

# Contents

# Abstract


## How can we ensure English language proficiency test scores are trustworthy and deserving of public confidence?

There are many challenges to maintaining test integrity in the global high-stakes assessment environment, and they are evolving with the growth of the test preparation industry, the proliferation of social media spaces, the rapid development of technology, and the demand for remote proctored testing. English language tests have the highest of stakes for test takers, who feel increasingly pressured to achieve top scores in order to access study or work opportunities. This pressure can unfortunately motivate some test takers to resort to shortcuts to achieve higher scores or test preparation behaviors that are not conducive to language learning. Such behaviors not only threaten the credibility of the assessment but also undermine the validity of the scores awarded. Testing organizations must constantly innovate to address rapidly evolving behaviors so that test scores carry meaning and can be trusted.

This paper discusses the current challenges to test integrity for high-stakes English language tests and the innovative solutions Pearson has implemented for the PTE Academic test to preserve test integrity through the strategic integration of AI and human judgment. While today's challenges far outstrip what human experts alone can handle, human judgment continues to play an irreplaceable role in every stage of the assessment, supported by AI systems.


> **We want to ensure that active human expert judgment is at the heart of our systems, combining the best of what automated systems can offer with the reassurance of nuanced human judgment.**

# Introduction

**The concept of test integrity is essentially concerned with trust – how *trustworthy* and how *trusted* a test is by its intended users. While validity remains the central concept underpinning the development, operation, and evaluation of high-stakes tests, it is vital also to consider the perceptions of stakeholders who decide which tests to trust and use in high-stakes decision-making.**

Stakeholder and public confidence and trust in high-stakes assessments may have been assumed in the past. For many years, the status quo of paper-based testing was a familiar environment for most test takers. This has gradually been replaced by online marking systems, computer-based testing and the advent of technology enhanced testing, including the use of AI. All of these advances have resulted in faster, more secure and reliable assessments. However, as these enabling technologies become more sophisticated, we need to ensure that all stakeholders are confident that the best that technology can offer us is as good as the best that expert human judgment can offer, albeit faster, and with far less error or unintended bias.

The operation and logistics of large-scale testing are now rightly more open to interrogation and challenge from government bodies, regulators, educational institutions, and test takers. Therefore, although Pearson is at the forefront of utilising new technologies to score and to maintain test integrity, we also want to ensure that active human expert judgment is at the heart of our systems, combining the best of what automated systems can offer with the reassurance of nuanced human judgment.

**While today's challenges far outstrip what human experts alone can handle, human judgment continues to play an irreplaceable role in every stage of the assessment, supported by AI systems.**

This paper outlines the research and development that underpins how we use technology and expert human judgment in scoring the PTE Academic test in order to maintain test integrity, and gives insights into further research and development steps.

## Test preparation and washback

An important aspect of maintaining test integrity is understanding how test takers prepare for the test and how this impacts language teaching and learning. When high-stakes tests control access to opportunities, both teachers and test takers are motivated to prioritize success on the test over other language learning goals, and this may in turn impact teaching, learning, and test preparation practices (Green, 2013). The term "washback" encompasses the interplay between teaching and learning practices, the design of the test, and the policy context motivating test takers to succeed. Washback can be positive, neutral, or negative, depending on how well aligned teaching and learning practices are with test preparation practices.

Negative washback occurs when teaching and learning become focussed solely on success on the test to the detriment of teaching and learning practices designed to develop true language proficiency (Hughes, 1989). Test developers intend for tests to have a neutral or positive washback, meaning that preparation for the test accords with or improves teaching and learning practices. However, some variables influencing washback, such as unofficial test preparation materials or the pressures facing test takers, are beyond the remit of test developers. While test developers cannot control all variables influencing washback, test developers have a responsibility to understand a test's washback effect and to respond to changing behaviors in the testing population to ensure that the test remains a valid measure of language proficiency.

Pearson has long been committed to understanding test preparation and washback effects. As PTE Academic grew in popularity and was increasingly used in high-stakes decisions, unofficial test preparation resources became more widespread. While many of the resources encouraged construct-relevant preparation, some less relevant strategies became more prevalent. Pearson responded by carrying out our own research internally and by funding academic research grants to investigate the evolving nature of preparation activities and their impact on scores (Fan et al., 2021; Knoch et al., 2020). These external studies noted the emergence of specific test preparation behaviors, indicating that the evolving stakes and policy of context of PTE were motivating changes in test-taking behaviors that required further research:

**It seems, on the one hand, that the perceived importance and difficulty of the PTE Academic has generated intense washback as Green's model predicts and might also be expected for all tests used to meet the highly stringent permanent residency requirements in the Australian context. In addition, the level of challenge imposed by the current policy encourages learners to adopt whatever strategy they perceive to be effective, regardless of the benefits for language learning.**

– (Knoch et al., 2020)

Pearson encourages test takers to seek PTE Academic preparation activities that build their language proficiency in ways that will support their ability to communicate in academic institutions, skilled professions, and everyday life. However, increasingly, both our own research and independent research have shown that some test preparation activities are oriented more toward achieving a high score on the test than building meaningful language proficiency. The most aggressive test preparation activities of this kind seek to misrepresent a test taker's proficiency. Such attempts to "game" or "crack" the test are considered construct-irrelevant response strategies, meaning that the strategy the test taker uses in their response bypasses or obscures the skill the test was intended to measure (Bejar et al., 2014; Messick, 1996). Ultimately, test scores lose their meaning when they can be achieved by means other than those we are intending to measure, and this is a threat to both validity and test integrity.

## Construct-irrelevant response strategies (gaming)

Gaming is defined as the use of construct-irrelevant response strategies that misrepresent or obscure a test taker's true ability. Gaming can be conceived of as a spectrum of behaviors characterized by varying degrees of construct irrelevance, misrepresentation of ability, and test taker intent. We use the term "gaming" rather than "cheating" or "malpractice" to describe these behaviors because test taker intent and misrepresentation vary significantly depending on the behavior.

**"Since PTE Academic's launch in 2009, the automated scoring system has always identified off-topic and irrelevant responses through robust scoring rules and the use of a content gatekeeper."**

**While test developers cannot control all variables influencing washback, test developers have a responsibility to understand a test's washback effect and to respond to changing behaviors in the testing population to ensure that the test remains a valid measure of language proficiency.**

For example, some test takers have been coached to only give partial responses to question types such as Repeat Sentence, because they believe that a full response is not required and only presents additional opportunities for error. Conversely, some have been coached to fill all the response time so there is no silence, so they repeat phrases or pad out their response with irrelevant material. Additionally, some test takers have been advised that the scoring system prefers certain manners of speaking, so they attempt to emulate a British accent, speak robotically, or speak unnaturally quickly. In these examples, test takers do not intend to cheat the system. Rather, they have been ill-advised based on incorrect assumptions about the automated scoring system. Regardless of whether this advice effectively increases scores, it has a real impact on test takers and the way they prepare for and sit the PTE Academic test.

This report focuses on a gaming strategy referred to as *memorized pre-scripted response*. It is by far the most prevalent gaming strategy attempted, and it is the strategy targeted by recent enhancements to the automated scoring systems for one writing item type (*Write Essay*) and two speaking item types (*Describe Image* and *Lecture Retell*).

For the *Write Essay* item type, test takers write an argumentative essay of 200–300 words in response to a prompt. Test takers attempt to game this item type by memorizing pre-scripted responses and then writing them verbatim in the test, adding little to no original content. The introductory paragraph of an example template is shown below.

*In this burgeoning epoch of science and technology* ***TOPIC*** *have become an integral part of the rising debate. It can be strongly agreed upon the fact that* ***TOPIC*** *has some persuasive/contentious arguments in favor of it, however, there has always been some contestation about it. This essay will elaborate how* ***TOPIC*** *and how* ***TOPIC****, which will result in a well-supported conclusion.*

▶ *Example of a memorized template for Write Essay items.*

**While more rudimentary systems may rely on a simple count of words matching known templates, PTE Academic's gaming detection system has been designed to consider a number of feature measurements that quantify the similarity of the response to known templates, the amount of authentic content present, the density of templated content, and the coherence of the response.**

This essay template is not specific to PTE – it is widely available on numerous test preparation websites for a diverse range of high-stakes tests that use an essay item type. This template is highly determined, meaning that test takers provide almost no original content when using this template. Other templates exist that provide scope for more original content, but any template that relies on long passages of memorized pre-scripted material is considered gaming.

For *Describe Image*, test takers have 25 seconds to look at a chart/graph/diagram/picture and 40 seconds to describe what they see. For *Retell Lecture*, test takers hear an audio recording of up to 90 seconds. They are given 10 seconds to prepare and then must summarize what they have heard in 40 seconds. Similar to *Write Essay*, both speaking item types are subject to gaming techniques in which test takers memorize pre-scripted responses and then repeat them verbatim in the test, adding little to no original content. However, the templates used for these speaking item types differ from the templates used for the writing item type. While the writing item type often sees templates that are fully formed responses memorized in their entirety, the speaking item types often see more fragmented and repetitive templates combined with fully memorized introductory and concluding sentences. The example template below is for a Describe Image item type.

*I have a beautiful picture in front of me. I have 40 seconds to talk about this picture. Let me have a closer look. Upon having a closer look, I can see colors, shapes, and numbers. I can see* ***COLOR****. I can see* ***COLOR****. I can see* ***COLOR****. I can see* ***NUMBER****. I can see* ***NUMBER****. I can see* ***NUMBER****. Overall, the picture is very informative.*

▶ *Example of a memorized template for Describe Image items.*

As in the example of the essay template, this template can be found in use across high-stakes tests that use an image description task. In some cases, test takers even plan which colors and numbers they will include in their response, regardless of whether they appear in the image they are describing. This type of repetitive template is inappropriate because it relies heavily on memorized material, but it is also worth noting that the resulting response is not particularly sophisticated or high scoring.

Since PTE Academic's launch in 2009, the automated scoring system has always identified off-topic and irrelevant responses through robust scoring rules and the use of a content gatekeeper (Pearson, 2023). The "content gatekeeper" is shorthand that refers to the various scoring rules that require test takers to meet content requirements of an item type in order to receive a score for the item. If a response is inadequate length, completely off-topic, unintelligible, or in a language other than English, the content gatekeeper would be triggered, and the test taker would score zero for the response. The purpose of the content gatekeeper was to ensure that the content of a test taker's response was a true and authentic representation of their own ideas, and not the result of a gaming technique, such as the use of memorized templates. Early research indicated the automated scoring system was effective at identifying outlier responses, including off topic and irrelevant responses (Cheng & Shen, 2011; Lochbaum et al., 2013).

The original PTE Academic content gatekeeper was adept at identifying basic template use that produced responses that were irrelevant to the prompt. However, in recent years, templates have become more sophisticated. A single generic template, when read in isolation, may appear convincingly relevant to a specific topic, making such templates difficult for both the automated content gatekeeper and expert human raters to identify – a challenge facing all high-stakes English language testing.

While this challenge is common across high-stakes testing, there has not yet been a definitive solution to addressing it. Some testing organizations rely on human judgment alone to identify templates both in the test preparation space and during scoring. Other organizations have begun to develop automated flagging systems but have not yet demonstrated that they are robust enough to meet this challenge fully. Pearson's scoring enhancements represent a notable advancement in scoring technology.

> **"Ultimately, test scores lose their meaning when they can be achieved by means other than those we are intending to measure, and this is a threat to both validity and test integrity."**

# Scoring enhancements

In 2022, Pearson implemented enhancements to the automated scoring systems with the goal of disincentivizing test preparation strategies that rely on using memorized or pre-scripted responses. Now in 2024, Pearson has introduced an additional layer of human-in-the-loop quality assurance, leveraging the detection capabilities of the automated scoring enhancements alongside the judgment of expert human raters.

## Phase 1: Automated detection and autonomous score adjustment

The automated scoring enhancements introduced in Phase 1 of this project were the result of more than 5 years of research into test preparation strategies and test taker behaviors. Launched in 2022, the automated scoring enhancements were designed to detect gaming behaviors and autonomously adjust scores in response. This section reports on the design, development, and human validation work to support the implementation of Phase I.

### Design

A gaming detection system was designed for one writing item type (*Write Essay*) and two speaking item types (*Describe Image* and *Lecture Retell*). Automated scoring of extended spoken and written responses is carried out by different scoring engines within the automated scoring system (Pearson, 2019). Consequently, separate gaming detection systems were developed for the speaking and writing scoring engines. The detection systems were developed in close consultation with each other to ensure consistency of scoring logic and standards.

**Human validation is necessary to implement automated scoring enhancements fairly and responsibly. It is vitally important that the automated scoring system does not act with bias or disadvantage particular test takers.**

Though the gaming strategies differ slightly for writing and speaking, common principles guided the design of both detection systems. Both writing and speaking detection systems were designed so that each response is evaluated for gaming. Based on the relationship between the response and known templates, the system calculates a "gaming score" for each response that represents the likelihood that the response contains a significant amount of memorized templated material. Both systems are based on machine learning algorithms that have been trained to use a suite of feature measurements to predict human ratings of gaming so that the higher the gaming score, the more likely a human rater would flag a response for gaming.

The gaming score is reported on a scale from 0 to 1, with 0 indicating no evidence of gaming and 1 indicating significant evidence of gaming. This reflects the reality that gaming is a matter of degrees rather than the absolute state of being gamed or not gamed. For example, we expect a certain amount of commonality across argumentative essays. In fact, a complete absence of common phrases used to scaffold and organize writing (e.g., "in conclusion" or "on the contrary") would be to the detriment of the essay's structure and coherence. While some common phrases are expected, significant passages of unoriginal text are not acceptable. However, there is a gray area in which test takers may employ some unoriginal phrases found in the repository of known gaming templates, but still go on to produce a fully developed, authentic response to the prompt. This gray area is notoriously difficult for even expert human raters to navigate, and the gaming score offers a standard frame of reference for judging responses in this range.

Importantly, the gaming score is based on multiple dimensions that take into account the nuanced nature of spoken and written communication. While more rudimentary systems may rely on a simple count of words matching known templates, PTE Academic's gaming detection system has been designed to consider a number of

**Any template that relies on long passages of memorized pre-scripted material is considered gaming.**

> **Gaming behavior is not naturally binary but requires individual human raters to consider whether a line has been crossed within a gray area.**

feature measurements that quantify the similarity of the response to known templates, the amount of authentic content present, the density of templated content, and the coherence of the response.

The strength of both the writing and speaking detection systems is that they provide standardized measurements of gaming as a spectrum of behavior and provide outputs to aid in explainability. This ensures the systems are flexible enough to adapt to changing test taker behaviors and policy demands. This also provides control over the severity, leniency, and meaning of gaming decisions because all decisions are referenced to a common, interpretable scale of gaming scores.

In Phase I, the gaming score was designed to be used by the automated scoring system to autonomously adjust scores. Responses with scores over a given threshold received a score of 0. Importantly, this scoring enhancement did not represent a change in scoring criteria, rather, it enabled the existing scoring criteria to be applied with increased precision to the evolving patterns of test taker responses.

**Development**

There were two phases to the development of each system: first, constructing a template repository of all known instances of templates for each item type, and second, building and training models to predict human ratings of gaming.

For the writing system, the template repository was constructed using a machine learning approach in which a large number of test taker responses were compared, their level of similarity assessed, and common templates and sub-templates extracted. This approach was particularly useful for writing, where templates tend to be longer and more varied, leading test takers to mix-and-match sub-templates as needed. The template repository was reviewed by human researchers to ensure that the templates and sub-templates included were composed of unique templated text, and not common phrases that would be expected to scaffold essays. Each template in the repository can be traced back to a test preparation website or test taker forum that encourages test takers to memorize the response in part or in full.

For the speaking system, the development grew out of an existing research program based on human labeling of gaming methods, which resulted in a human-constructed template repository. The template repositories were evaluated across several prompts to determine if test takers were memorizing responses to specific prompts, which would indicate that the items themselves had been compromized. However, it was found that

the same templates were common across a range of prompts, which indicates that test takers generally memorize prompt-agnostic templates, and then make varying degrees of effort to tailor them to the prompt they encounter in the test.

Human-labeled data was used to train each model, and thresholds were set to maximize the precision of automated detection.

**Human validation**

Once the gaming detection systems had been developed, they were evaluated through a human validation process. The human validation process was designed to assess whether the automated detection system aligned with the standard set by human raters and could be deployed in the live testing environment to autonomously adjust scores. In the live environment, responses would be flagged for score adjustment based on whether their gaming score exceeded a given threshold. The validation process tested whether human raters would agree that the scores should be adjusted for responses that exceeded this threshold. In normal operation, human raters judge a response to be gamed and adjust the score if two criteria are met: (1) a significant proportion of the response consists of memorized or pre-scripted text and (2) the authentic portion of the response does not adequately address the prompt or demonstrate a good faith attempt to answer the question.

Human validation is necessary to implement automated scoring enhancements fairly and responsibly. It is vitally important that the automated scoring system does not act with bias or disadvantage particular test takers. Accordingly, the success criteria for human validation were based on a conservative approach that prioritized defensibility, precision, and in particular, a low false positive rate. A "false positive" occurs when the machine detects gaming in a response, but the human judge does not. The goal was to minimize false positive results and ensure that any discrepancies between human and automated system did not show evidence of bias and could be explained. Alongside the quantitative analysis of model performance, qualitative analysis of all false positive responses was undertaken to ensure that no responses were automatically flagged for gaming without significant, defensible reason.

For the written detection system, the human validation sample included 2,000 randomly selected responses to *Write Essay*. The responses were rated for gaming by the automated detection system, human scorer 1 (HS1), and human scorer 2 (HS2). Human raters reviewed the

> **"Pearson has introduced an additional layer of human-in-the-loop quality assurance, leveraging the detection capabilities of the automated scoring enhancements alongside the judgment of expert human raters."**

responses independently and without knowledge of the gaming score. Of the 2,000 responses in the sample, the two human raters agreed on the rating for 1,530 responses. Of those responses, only 16 (1.04%) false positives were identified in which both human raters did not flag a response for gaming, but the automated system did. For the spoken detection system, the human validation sample included 1,000 responses. The responses were equally drawn from *Describe Image* items and *Retell Lecture* items. Each response was rated by the automated system, HS1, and HS2. Of the 1,000 responses, both human raters agreed on the classification of 823 responses. Of those responses, no false positives were identified. For every response the system flagged for gaming on speaking items, both humans agreed that it was in fact gamed.

A qualitative evaluation was carried out on each false positive response to ensure that the automated detection system's gaming flags could be understood and justified, even where human raters had not independently flagged the response for gaming. In all cases, the "false positive" responses contained significant amounts of text originating in known gaming templates, and the responses were ultimately determined to have been correctly identified by the automated system. No true false positives were found in the human validation of either the written or spoken detection system. A summary of outcomes for the human validation of both systems is shown in Table 1.

**Table 1. Summary of human validation outcomes**

| System | Human validation sample size | N human agreement | Human agreement rate | N false positives | N confirmed false positives after qualitative review |
|---|---|---|---|---|---|
| **Speaking** | 1,000 | 823 | 82.3% | 0 | 0 |
| **Writing** | 2,000 | 1,530 | 76.5% | 16 | 0 |

> **"The human validation process was designed to assess whether the automated detection system aligned with the standard set by human raters and could be deployed in the live testing environment to autonomously adjust scores."**

> **The system calculates a "gaming score" for each response that represents the likelihood the response contains a significant amount of memorized templated material.**

In fact, when considering the information in Table 1, the term "false positive" may be misleading in the context of this study. Because gaming decisions often fall within a gray area and reasonable expert human raters may disagree on a rating, it is impossible to allocate every response in the sample to either a "true positive" or "true negative" category. Table 1 shows that humans agree with each other on whether a response is gamed about 80% of the time. This highlights the point that gaming behavior is not naturally binary but requires individual human raters to consider whether a line has been crossed within a gray area. A shared, standardized concept can be developed, but it is unlikely to generate perfect agreement between human raters. This may be attributable, in part, to differing levels of leniency or severity. It may also be attributable to differing levels of familiarity with gaming templates. In some cases, humans may not be familiar with specific templates and unable to accurately flag responses that use them. Additionally, typical forms of human error are possible, such as mislabeling or misreading responses.

Ultimately, the human validation data indicates that the detection systems are able to detect gaming behavior in extended written and spoken responses and reliably flag responses for score adjustment when the response demonstrates sufficient gaming behavior that two human raters would independently agree to adjust the score.

## Phase 2: AI-enabled human oversight

The Phase I enhancements to the automated scoring system represented a substantial innovation in the ability to detect the use of memorized pre-scripted responses and adjust scores automatically to account for them. The limitation of the Phase I approach was in fact a human limitation. Because there is a limit to where humans will consistently agree on gaming decisions, there is also a limit to what the automated scoring system can be trained to do autonomously and responsibly. As a result, Phase I was designed to be conservative.

Phase II is designed as an extension to Phase I, enabling more fine-grained judgments by combining both human and automated scoring system judgments. The primary limitation of human raters in identifying gaming behaviors is their limited exposure to the range of templates in circulation at any one time. This Phase II development enables raters to make better informed content scoring decisions by leveraging the Phase I gaming detection systems to provide them with information about the extent of gaming behaviors detected in the response.

> **The primary limitation of human raters in identifying gaming behaviors is their limited exposure to the range of templates in circulation at any one time.**

### Design and development

In 2024, Pearson introduced Phase II, an additional layer of quality assurance in the scoring of the PTE Academic test. All responses will continue to be scored by the automated scoring systems and anomalous responses will continue to be scored by human raters. In addition, the content of *Describe Image* and *Retell Lecture* responses will now also be scored by human raters alongside the automated scoring system, supported by information from the Phase I detection system, as shown in Figure 1 below. If the human rater and automated detection system disagree on the content score, a senior human rater will adjudicate, and their score will be final.

For Phase II, gaming detection information for *Describe Image* and *Retell Lecture* responses will be passed to human raters to help inform their scoring decisions. The automated detection system produces a gaming score based on the combined measurements of several features related to gaming. This gaming score is summarized and passed to human raters in a way that describes the strength of the evidence found by the automated system, either "Significant evidence of gaming", "Some evidence of gaming", or "Little to no evidence of gaming". These category labels were developed in consultation with the team of human raters. The raters expressed a preference for clear and direct

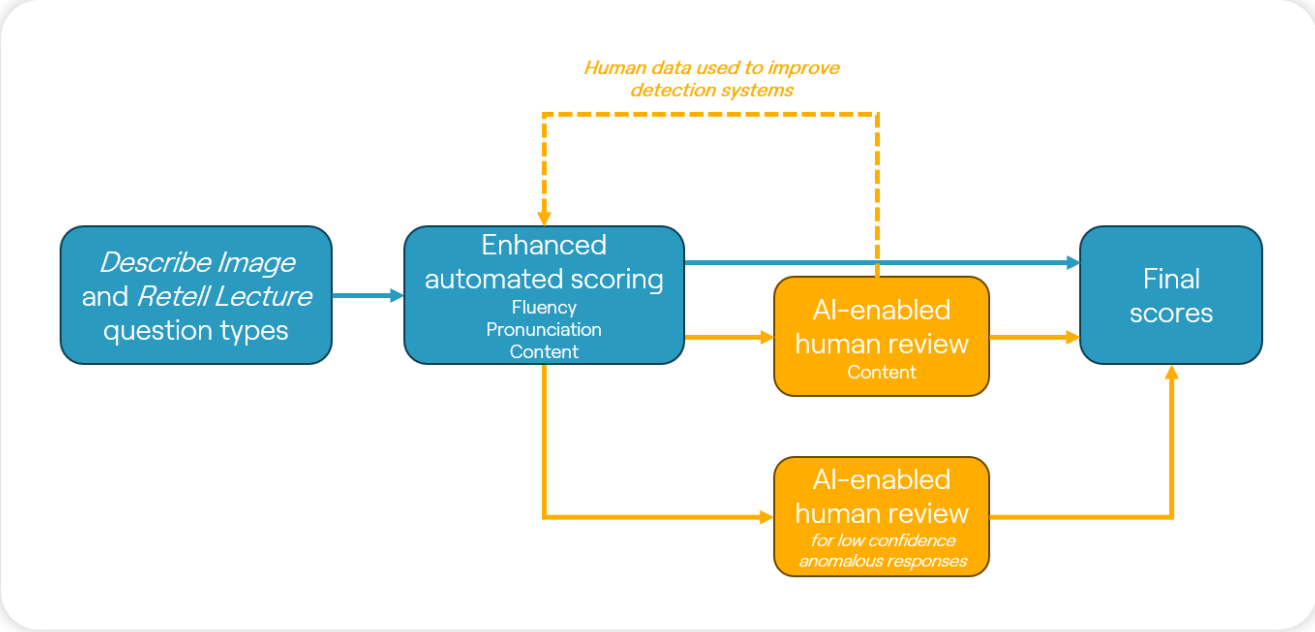


**Figure 1.** *Enhanced scoring process with additional layer of human oversight.*

labels rather than a detailed report of all gaming measurements. From launch, rater feedback will be essential to understand the utility of these labels and to explore alternative methods of presenting gaming information to human raters.

Where the Phase I system can only adjust scores when responses demonstrate high levels of gaming strategies, the Phase II system enables human raters to apply their judgment to gaming decisions while benefiting from the information produced by the automated system. In turn, the automated detection system will learn over time from the continuous stream of human rater judgments to become more sensitive to gaming behaviors.

While we think this makes for a better, more accurate scoring system, we also believe it makes for a more responsible one. We recognize that there are some decisions we are more comfortable with humans making, particularly in gray areas. In that case, the role of technology is to support human decisions to be as evidence-based and consistent as possible.

**Validation**

A Phase II validation study was run using a sample of 998 test takers taking part in a PTE field test in 2024. During the field test, raters were not able to view information from the automated detection system, but speaking scores were calculated using both the Phase I (automated detection only) and Phase II (automated detection and human scoring) scoring models.

The correlation between the speaking scores awarded under the Phase I scoring model and those awarded under the Phase II model was 0.95. For the test takers who are not relying on memorized responses, scores will not be impacted by this additional layer of quality assurance. Test takers who are relying excessively on memorized responses will likely also see little change, as their behavior is already detected by the current scoring system and their scores are adjusted appropriately. Test takers whose behavior has been under the high confidence threshold set for the automated system, but above the threshold of AI-enabled human judgment, will see their scores adjusted appropriately with the launch of Phase II.

> **“This gaming score is summarized and passed to human raters in a way that describes the strength of the evidence found by the automated system, either “Significant evidence of gaming”, “Some evidence of gaming”, or “Little to no evidence of gaming”.”**

# Monitoring impact and future work

**A robust program of monitoring and maintenance is needed to respond to rapidly evolving test taker behaviors. Figure 2 shows the development cycle for continuous monitoring and improvement of automated detection systems that has been in place since 2022.**

We monitor online spaces for the emergence of new templates and gaming strategies through web crawling and social media monitoring. We regularly undertake template discovery, which uses machine learning techniques to analyze very large samples of real test taker responses to identify new patterns of highly similar text. We deploy minor updates to the template repository to ensure it can identify new templates currently in circulation. Finally, we conduct research to investigate fairness and accuracy, and inform major updates to improve the underlying detection algorithms. In Phase II, the monitoring and improvement program will also benefit from continuous human validation data fed back into the detection systems to evaluate their alignment with human raters and improve the accuracy of detection models.

The automated scoring system enhancements described in this report have been effective and ultimately necessary to combat the negative washback and construct-irrelevant response strategies arising in the highly competitive test preparation industry. Pearson remains committed to continuously monitoring test taker behaviors for emerging gaming strategies, expanding automated detection to additional item types, and pioneering new automated detection methods.

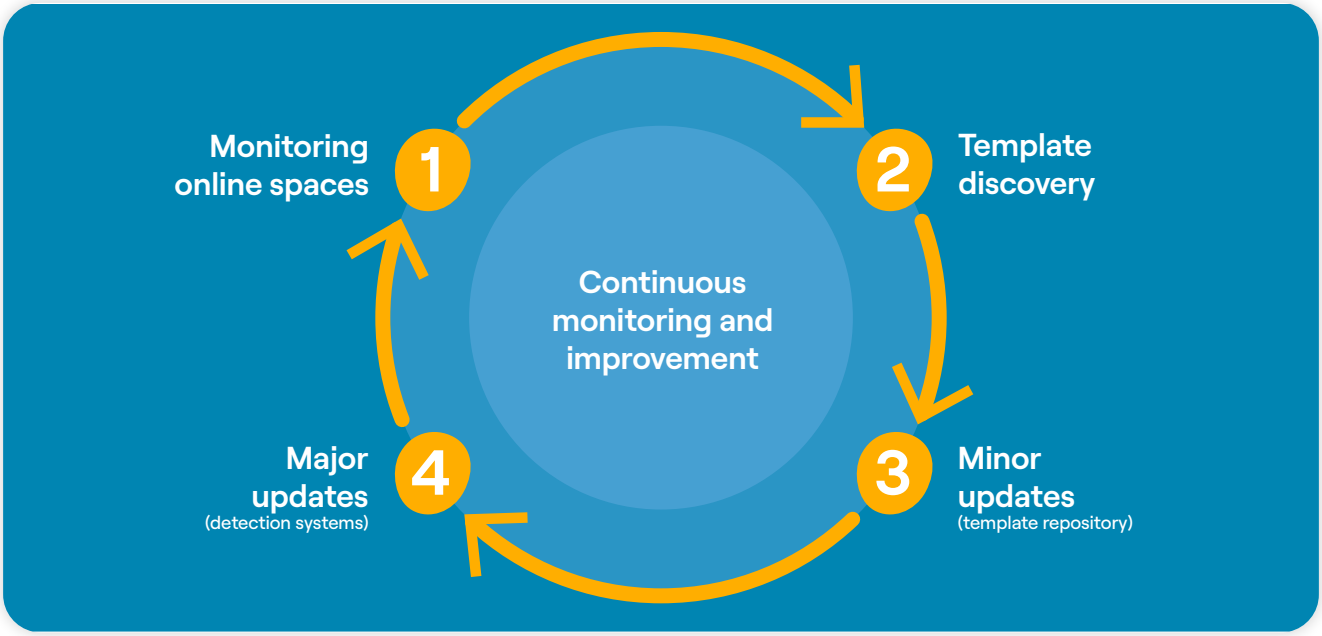


**Figure 2.** *Development cycle and continuous improvement of automated detection systems.*

> **In globally run assessments, it is not possible for human raters to develop an overview of response patterns for the whole testing population, nor is it possible for individual human raters to remain current with continuously evolving test taker strategies.**

The research supporting the development and validation of the automated gaming detection system has highlighted the complementary strengths of human and automated scoring. While human judgment is necessary to conceptualizing gaming behavior, the anti-gaming scoring enhancements described here far exceed the capabilities of human raters alone. In globally run assessments, it is not possible for human raters to develop an overview of response patterns for the whole testing population, nor is it possible for individual human raters to remain current with continuously evolving test taker strategies.

Finally, it should be noted that gaming detection is only one side of the coin. The other side is gaming mitigation. Future innovation should move toward designing item types that are more resilient to memorized or pre-scripted responses. Currently, all high-stakes English language tests use some form of essay task, and templates can be found for each of these tasks in online test preparation communities. Innovation is needed across the field of language assessment to design tasks that are resilient to gaming behaviors, and at the same time test the intended construct with authenticity and reliability.

However, even with the advent of new item types, assessment organizations should be prepared for highly motivated test takers to learn to game new item types with time.

As noted by Green (2013):

> **The imperative to succeed on a test encourages teachers and learners to adopt short-term strategies, prioritizing memorization of large amounts of content over building a deeper understanding of underlying principles. The most deleterious effects come from high-stakes tests that control access to opportunities and so are seen as very important to test takers' life chances. The choice of test format and content may have a relatively trivial impact on this behavior.**
>
> – Green (2013)

As long as the stakes are high and the requirements are demanding, there will be incentive to game that no amount of innovative assessment design can remove. In this context, assessment organizations have a responsibility to constantly monitor the behaviors of the test taking population and to adapt flexibly to maintain the integrity of test scores.

# Conclusions

**This paper has described the research and development involved with dealing with the ever-increasing threats to test integrity in high-stakes testing and at the same time ensuring that that there is confidence and transparency in the integration of sophisticated automated technologies alongside expert human judgment when identifying gaming behaviors and scoring extended speaking and writing responses.**

Most test takers prepare for and take high-stakes English language tests in good faith, and it is therefore important that test takers and stakeholders have trust in the operation and standards of the assessment system. As testing organizations, it is our responsibility to ensure that our tests are as valid, reliable, and as fair as possible and that we regularly monitor both the qualities of the tests themselves and any emergent behaviors that encourage negative washback in terms of test preparation and test taking strategies.

As outlined in this paper, the emergence of technology has transformed the delivery, accuracy, and fairness of global English language testing. It is an obvious corollary that the use of technology is also increasingly required to protect test integrity. The development of automated monitoring systems that can flag gamed speaking and writing responses which can then be considered by human expert raters utilises both the power and reach of technology, paired with the essential element of human judgment.

As well as dealing with the increasing threats of templating and negative washback behaviors, the duality of technology and human judgment can also provide a baseplate for evidencing the accuracy and reliability of Pearson's automated scoring systems and ensure that public trust and confidence in high-stakes assessment is evidenced and not taken for granted.

The maintenance and expansion of testing integrity, alongside reliability, validity and fairness measures will continue to be a focus of Pearson's ongoing research and development programs.

**"The development of automated monitoring systems that can flag gamed speaking and writing responses which can then be considered by human expert raters utilises both the power and reach of technology, paired with the essential element of human judgment."**

# References

Bejar, I., Flor, M., Futagi, Y., & Ramineni, C. (2014). On the vulnerability of automated scoring to construct-irrelevant response strategies (CIRS): An illustration. *Assessing Writing, 22,* 48–59. https://doi.org/10.1016/j.asw.2014.06.001

Cheng, J., & Shen, J. (2011). *Off-topic detection in automated speech assessment applications.* Twelfth Annual Conference of the International Speech Communication Association.

Fan, J., Jin, Y., Kong, X., & Zhang, X. (2021). *How do test takers prepare for computerized speaking tests? A comparative study of the PTE Academic and the CET-SET.* Language Testing Research Centre, University of Melbourne.

Green, A. (2013). Washback in language assessment. *International Journal of English Studies,* 13(2), Article 2. https://doi.org/10.6018/ijes.13.2.185891

Hughes, A. (1989). *Testing for language teachers.* Cambridge university press.

Klungtvedt, M., & Van Moere, A. (2012). *Automatic scoring of off-topic and marginal responses*. Pearson (internal report).

Knoch, U., Huisman, A., Elder, C., Kong, X., & McKenna, A. (2020). Drawing on repeat test takers to study test preparation practices and their links to score gains. *Language Testing,* 37(4), 550–572.

Lochbaum, K., Rosenstein, M., Foltz, P., & Derr, M. (2013). *Detection of gaming in automated scoring of essays with the IEA*. Paper presented at the National Council on Measurement in Education (NCME) Conference, San Francisco, CA.

Messick, S. (1996). Validity and washback in language testing. *Language Testing,* 13(3), 241–256.

Pearson. (2019). *PTE Academic: Automated Scoring*. https://assets.ctfassets.net/yqwtwibiobs4/26s58z1YI9J4oRtv0qo3mo/88121f3d60b5f4bc2e5d175974d52951/Pearson-Test-of-English-Academic-Automated-Scoring-White-Paper-May-2018.pdf

Pearson. (2023). *PTE Academic Score Guide*. https://assets.ctfassets.net/yqwtwibiobs4/3Bm0RMkKoNVOoOxUe38mg4/f565a92a97e8f3cf60c5506d347dedb8/PTE_Academic_Score_Guide_for_Test_Takers_June_2023.pdf

Pearson

## About the authors

### Dr. Rose Clesham

Dr. Rose Clesham is the Director of Assessment Research and Validity. She holds a Master's Degree in Formative and Summative Assessment from Cambridge University and a Doctorate in Educational Assessment. She is a Fellow of the Association for Educational Assessment-Europe (AEA-E) and an Honorary Associate Professor at the University of London (UCL)

She has worked extensively on OECD PISA assessments, including co-writing assessment Frameworks. Rose lectures on international educational standards, validity and reliability issues. Her research interests include the development of e-assessment and Artificial Intelligence systems, and on-going international educational strategies and policy.

### Sarah Hughes

Sarah is Head of Test Design and Validation in Global Assessment at Pearson. Working in the Global Product English Language Assessment team, with a focus on research into Pearson Test of English - Academic (PTE-A) and large-scale high-stakes language testing.

Sarah has a Master's in English from Brown University. Alongside her work at Pearson she is a PhD candidate in the Cambridge Faculty of Education, researching validity and AI in assessment.

Pearson | PTE

Secure. Accurate. Trusted.

pearsonpte.com/research