

Responsible AI in the Duolingo English Test: Case Studies with Automatic Item Creation and Session-Level Quality Monitoring

Siyuan Marco Chen Xiaowan Zhang Andrew Runge Jacqueline Church

Duolingo, Inc.

Pittsburgh, PA, USA

{marco, xiaowan, arunge, jacqueline}@duolingo.com

Abstract

Digital-first assessments are delivered continuously and often remotely. Artificial intelligence (AI) enables digital assessments to generate content and administer tests at scale. Like any system used for high-stakes decision-making, digital assessments require responsible AI (RAI) practices to ensure fairness and validity. This paper presents two deployed systems in the Duolingo English Test (DET) lifecycle that align the DET to its RAI Standards. The Item Factory combines automated item generation with staged expert review; the Analytics for Quality Assurance in Test Taker (AQUA-TT) system applies unsupervised anomaly detection methods to continuously monitor for issues in digital test deliveries for daily individual test sessions. We present the design and performance of these systems and discuss what they imply for placing human judgment inside digital assessments.

1 Introduction

Digital-first assessments (Langenfeld et al., 2022) are delivered continuously and at scale to test takers, in contrast to traditional in-person assessments that are administered on scheduled dates and locations. Digital-first assessments require automated content generation and continuous score production, which often depends on artificial intelligence (AI) systems, to meet the demand of large-scale item exposure and test results (Liao et al., 2022). Such automated processes allow for no static time windows before test delivery for one-off content reviews or continuous human proctors during every test administration. In other words, having only fixed and non-continuous monitoring is insufficient for AI-enabled digital assessments and can more easily lead to risks, such as item leakage or unintended test delivery errors. These in turn can affect the validity of scores and fairness of high-stake decisions (Fiesler and Garrett, 2020). As a

result, responsible AI (RAI) practice guidelines (e.g., Burstein et al., 2025), which defines practices to reduce vulnerabilities and fairness concerns in AI systems, should be followed across all stages of digital test development and administration (von Davier et al., 2026).

One example of following RAI guidelines is the Duolingo English Test (DET) (Naismith et al., 2026), a high-stakes, digital-first English proficiency test used for higher-education admissions. This test aligns with The Duolingo English Test RAI Standards (Burstein, 2026), which incorporates foundational validity and fairness frameworks to the digital assessment context and rests on four standards: Validity and Reliability, Fairness, Privacy and Security, and Accountability and Transparency. The DET RAI Standards extend beyond the Standards for Educational and Psychological Testing (AERA, APA, and NCME, 2014) to specify DET practices in automated item generation, scoring, as well as test administration. In particular, the DET RAI Standards emphasize that human oversight is essential as a form of quality assurance at critical stages of assessment development and evaluation.

The DET operationalizes RAI practices by developing innovative human-in-the-loop AI systems for the development of item bank, score production, and the monitoring of test delivery (von Davier et al., 2026). This paper presents two deployed instantiations of such systems, the Item Factory and Analytics for Quality Assurance in Test Takers (AQUA-TT). The Item Factory pairs automated generation with staged expert review before items reach the operational bank; AQUA-TT conducts automated and continuous monitoring of individual test sessions, applying unsupervised anomaly detection in the test delivery process to facilitate efficient human review. These two systems together embody all four standards of RAI. They demonstrate how (a) human-in-the-loop automated test develop-

ment processes and (b) innovative anomaly detection methods implement RAI standards at scale in support of the fairness and validity of digital-first assessment outcomes.

2 Case Study 1: Item Factory for Human-in-the-Loop Automatic Item Creation

2.1 Automatic Item Generation (AIG) Workflow

The Item Factory is a system that automates the production of test items for the DET. This infrastructure supports the creation of high-quality test content through responsible human–AI collaboration (von Davier et al., 2024). Test tasks and their components are developed and improved through a four-stage iterative process: (1) building an AIG system, (2) conducting quality and fairness/bias reviews, (3) collecting expert feedback on validity and system performance, and (4) incorporating feedback into subsequent system iterations. Below, we illustrate this workflow through a study aimed at improving the item generation pipeline for the DET Interactive Reading task, in which test takers answer a series of comprehension questions based on a short passage.

2.2 Methods

The Interactive Reading task generation study was conducted over a three-month period in 2023. Four batches of tasks were generated, with each task including either an academic or a nonacademic narrative passage and a set of comprehension questions. Next, we briefly describe the processes for generating the materials as well as how we iterated on the generation pipelines across batches.

2.2.1 Passage Generation

Passages were generated using GPT-4 through a multi-step, compositional process following Attali et al. (2022). Prior to generating passages, we first generated content outlines through a series of branching steps in which we iteratively prompted the LLM to expand a broad academic topic or general story idea with more specific details. For narrative passages, these details included setting and character information, and for academic passages, they included important terms, events or discoveries. We then used another LLM prompt to generate passages based on the outlines. This second prompt took in additional pre-defined passage specifica-

tions, such as a target audience (e.g. middle school students), a style of writing (e.g. fun and engaging), and a target length in order to produce greater variability in the language of the final passages. Each round of generation typically yielded hundreds or thousands of passages. For this study, we randomly selected a sample of the generated passages for human review, balancing for topic and passage length (150–225 words). The four batches contained 80, 135, 135, and 206 passages, respectively, and their associated items.

2.2.2 Distractor Generation

Although the reading task includes several question types, this paper focuses on the main idea question (selecting the option that best describes the passage’s main idea) due to space constraints. In Batches 1–2, distractors for this question type were generated using an alternative-passage strategy, in which multiple versions of a passage were produced from similar inputs and used as sources for distractor options (Attali et al., 2022). In Batches 3–4, we compared the alternative-passage method to a new direct-generation approach. This approach generated distractors directly from the target passage and correct answer by prompting the model to adopt different strategies for misinterpreting information in the passage. For each question, a pool of candidate distractors was generated, from which up to six were automatically selected for human review using several metrics, including the LLM-estimated likelihood of each answer (see details in Attali et al., 2022). To examine how reviewers responded to the new method, the automated selection process in Batch 3 was constrained so that, among the six distractors selected for human review, no more than one was generated using the direct approach, while the remainder were generated using the alternative-passage method. We removed this restriction in Batch 4 and simply auto-selected the top six distractors for review based on metrics, regardless of their generation origin.

2.2.3 Item Review

Twelve experienced item reviewers evaluated and edited the automatically generated passages and questions following internal guidelines. They edited the passages to improve their fluency, logical coherence, and appropriateness. Reviewers were instructed to fail a passage if it required extensive and time-consuming edits. Reasons for rejection included (a) excessive jargon, (b) inappropriate or

sensitive topics, (c) low content accessibility (e.g., culturally specific content), (d) implausible logic, or (e) stereotypical representations. In addition, reviewers edited the questions to ensure that keys were accurate and that distractors were attractive and non-keyable. For each question, they selected four of the six distractors presented for human review and, when necessary, edited them for operational use.

2.2.4 Item Generation Iteration and Evaluation

Reviewer comments, task failure tags, and reviewer editing behavior were all used as feedback to iteratively refine the generation pipeline across the four rounds of item generation. Based on feedback from the first batch, the second batch focused on making improvements to passage generation, whereas the latter two focused on improving comprehension-question generation, particularly the generation of multiple-choice distractors. Passages generated in Batches 1 and 2 were compared to evaluate changes in quality using several metrics, including the percentage of passages that passed with edits, the extent of reviewer editing (measured by edit distance), and the distribution of failure tags. For distractors generated in Batches 3 and 4, the two generation methods (direct vs. alternative-passage) were compared in terms of edit distance and the number of distractors ultimately selected by reviewers. Following operational administration, the methods were also compared based on the attractiveness and discrimination of the selected distractors.

2.3 Results

2.3.1 Passage Generation

For Batch 1 and 2, we focused on iterating on the passage generation to improve the passage AIG pipeline. In Batch 1, 77.5% of passages (62/80) passed with edits. The most common issues with this batch were excessive jargon and inappropriate topics. Reviewers also noted predictable passage structures and styles, including the overuse of certain words (e.g., “newfound”). Based on reviewers’ feedback, the passage AIG system was refined by (a) updating prompts to require explanations of potentially unfamiliar or technical jargon, (b) applying stricter GPT-based filtering of topics, and (c) adding style descriptors to increase lexical and grammatical diversity. Changes to the passage generation pipeline led to clear improvements in Batch 2. The pass-with-edit rate increased to

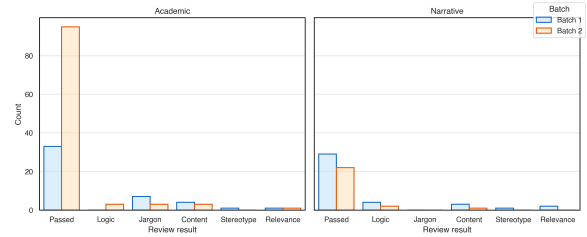


Figure 1: Distribution of Failure Tags Across Passages Generated in Batches 1 and 2 by Passage Type

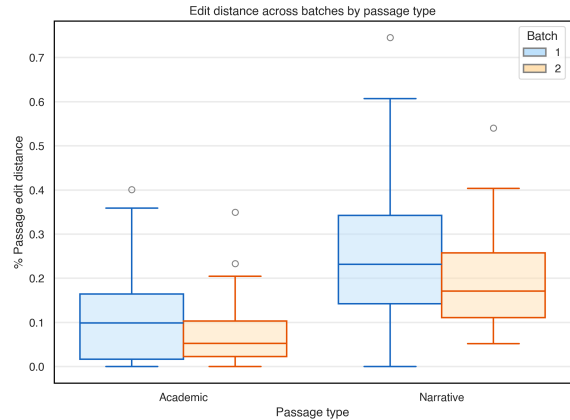


Figure 2: Normalized Character Edit Distance for Passages Generated in Batches 1 and 2 by Passage Type

86.7% (117/135), while failures related to jargon and topic suitability decreased substantially (see Figure 1 and Table A1 in Appendix A). Editing effort also decreased. As illustrated in Figure 2, the normalized character edit distance decreased by 6.9% for academic passages and 11.8% for narrative passages (see Table A2 in Appendix A). Mixed-effects model with batch as a fixed-effect variable and reviewer as a random-effect variable confirmed significantly lower overall edit distance in Batch 2 ($p < .001$; see Table A2 in Appendix A), consistent with reviewers’ qualitative feedback. However, not all changes were effective. For instance, introducing individual style variations did not yield meaningful gains, and some stylistic prompts resulted in overly frequent, less-desirable patterns (e.g., similar openings beginning with “Have you ever wondered. . .”).

2.3.2 Distractor Generation

For Batch 3 and 4, we focused on improving distractor generation for MCQs. We next report the results for the main-idea question. The alternative-passage method produced highly varied distractors, but they often required extensive reviewer revisions. In contrast, the direct approach produced more contex-

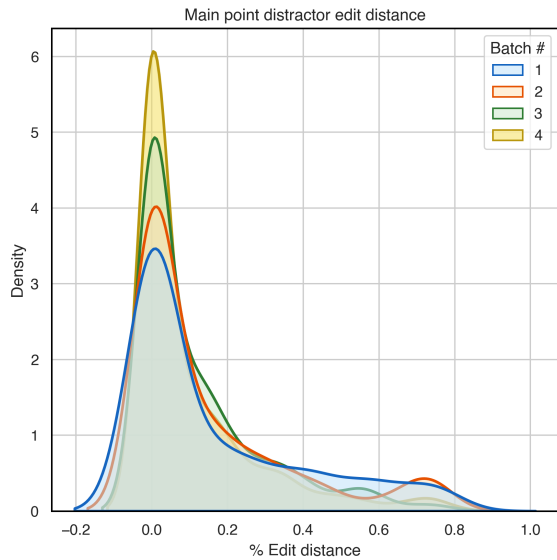


Figure 3: Density Plot of Normalized Character Edit Distance for Main-Idea Question Distractors Across Batches 1–4

tually relevant distractors, although their number was more limited. Figure 3 shows the normalized character edit distance of the distractors across the four batches. Edit distance decreased slightly from Batch 1 to Batch 2, likely because higher-quality passages required fewer downstream revisions to questions and distractors. Larger improvements were observed in Batches 3 and 4, in which directly generated distractors were introduced.

Among distractors presented for human review, directly generated distractors were retained more frequently than alternative-passage distractors in both Batch 3 (76.25% vs. 67.30%) and Batch 4 (76.36% vs. 64.08%). Among those retained for operational use, directly generated distractors also required less editing in both batches. Mixed-effects model with method as a fixed-effect variable indicated that the difference in editing was statistically significant in Batch 4 (see Table A3 in Appendix A).

In addition, pilot response data showed that directly generated distractors were more attractive to test takers, as indicated by a higher average selection rate (8.61% vs. 6.35%, $p < .001$; see Figure 4 and Table A4 in Appendix A). Moreover, they exhibited slightly stronger discrimination, with more negative point-biserial correlations ($-.25$ vs. $-.23$, $p < .001$; see Table A4 in Appendix A). Together, these findings suggest that direct generation produces more plausible and effective distractors. Nevertheless, the two methods could be used comple-

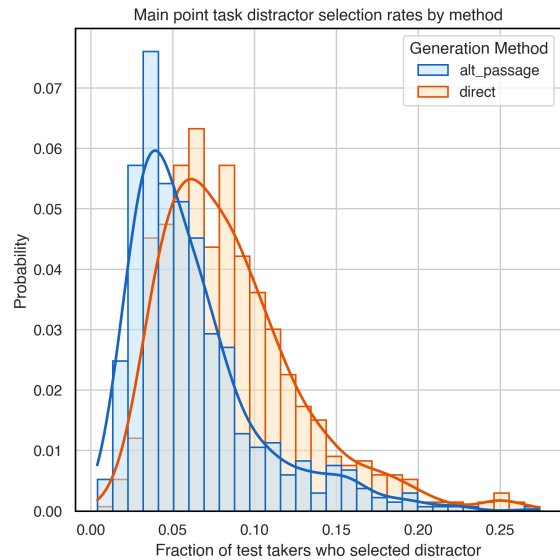


Figure 4: Distractor Selection Rate by Generation Method (Overall across Batches 3 and 4)

mentarily, with the alternative-passage method providing greater variety and the direct approach producing more contextually relevant options.

The Item Factory operationalizes three goals in the DET RAI Standards (Burstein, 2026): Goals 1.1, 1.2, and 2.1. Under Validity and Reliability, Goal 1.1 requires items to align with a defined target domain. The generation pipeline for the Interactive Reading task supports this alignment by incorporating the domain specifications directly into the generation process. Goal 1.2 calls for AI to leverage “human expertise to efficiently create valid and reliable test items,” including through “human evaluations of the quality of items created using AI” (Burstein, 2026, p. 9). The four-batch iterative process described in this study reflects this goal: expert reviewers evaluated all passages and questions using established guidelines, documented their edits and pass/fail decisions, and provided feedback that informed successive refinements to the generation pipeline. Under Fairness, Goal 2.1 requires the development and application of principles for fairness and bias review. Expert reviewers addressed this goal by applying the DET’s established fairness guidelines (Church et al., 2025) when reviewing the generated Interactive Reading tasks.

3 Case Study 2: Analytics for Quality Assurance in Test Takers (AQUA-TT)

The DET is administered continuously and remotely using automated item administration and scoring algorithms. Irregularities in the test delivery environment or unusual test-taking behaviors may alter how items are administered or how responses are recorded. Such irregularities may not produce a detectable shift in aggregated score distributions unless many sessions have accumulated over time (Liao et al., 2022). Existing session-level monitoring works in digital assessment have focused on detecting test-taker misconduct (e.g., Niu et al., 2025; Belzak et al., 2025). On the other hand, less attention has been given to detecting irregularities in test delivery, where the digital test platform itself shows unexpected issues (e.g., the software interface having a prolonged lag in loading items). We developed AQUA-TT to monitor incoming DET sessions and promptly surface sessions with risks of irregular item response patterns for human review, supporting the Validity, Reliability, and Accountability and Transparency principles of the DET’s RAI framework. AQUA-TT’s operational design goal is to enable human reviews on sessions affected by test delivery irregularities (e.g., bugs in the user interface) among the top ranked sessions that surfaced as risky on a given day, with acceptable alert precision based on historical data evidence. This helps to limit manual review effort across item types. Human review determines whether the surfaced pattern reflects benign test-taking behaviors or an underlying delivery issue warranting escalation.

3.1 Modeling Method and Features

We applied Gaussian Mixture Models (GMMs) to learn the multivariate distribution of normal response patterns for each of the DET’s item types, trained on a reference dataset of sessions representative of normal test-taking behavior. This approach builds a probabilistic model of normal behaviors within each item type of the DET, so that observations with low likelihood under a trained GMM are treated as anomalous (Chandola et al., 2009). Incoming DET sessions are assigned risk scores based on their negative log-likelihood (NLL) under each trained model. Sessions whose patterns of feature values for an item type appear atypical receive a high NLL and become candidates for review. The approach requires no labeled examples

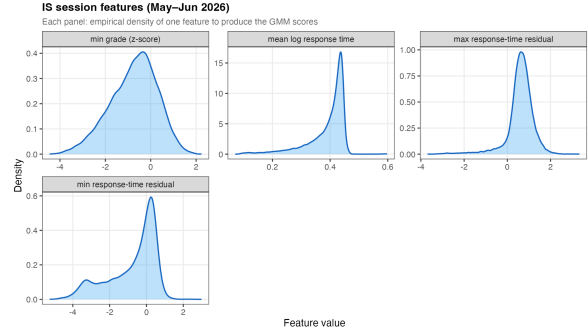


Figure 5: Empirical feature densities for the Interactive Speaking (IS) item type during May-June 2026 DET sessions.

of the irregularities being detected, which better generalizes to the detection of previously unknown delivery issues that have no training data available for supervised learning.

Features are engineered per item type to describe response patterns and serve as input to the GMMs. They include mean item grades, response time, deviations of observed response time from the time predicted by item difficulty and proficiency (Meijer and van Krimpen-Stoop, 2009), and a cumulative sum (CUSUM) psychometric statistic. The CUSUM person-fit statistics indicate the magnitude of aberrant patterns in observed grades relative to the response expected by the item response models (van Krimpen-Stoop and Meijer, 2002; Tendeiro and Meijer, 2012) and are computed for the adaptive item types.

Each GMM selects its numbers of component and covariance structures to approximate the multivariate feature using Integrated Complete-data Likelihood (ICL) criterion (Biernacki et al., 2000), which better reduces overfitting to the training data relative to Bayesian Information Criterion (BIC). See Appendix B for details on ICL. GMM models are fit using the `mclust` R package (Scrucca et al., 2016). Figure 5 shows examples of these feature densities for our Interactive Speaking item type on recent sessions. They serve as input for the GMM to learn the joint density of this item type, which is shown in Figure 6. Next, we demonstrate an alert system that operationalizes this detection method into an automated monitoring tool.

3.2 Alert System Design

AQUA-TT employs an alert system to (1) surface sessions with high risk of having abnormal response patterns (due to test-taking behaviors or delivery issues) and (2) notify human reviewers to

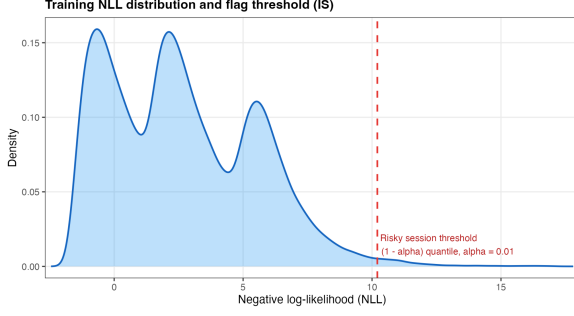


Figure 6: Distribution of IS risk scores (negative log-likelihood under the IS GMM) across reference training sessions. The dashed line marks the session-level flag threshold, set at the $(1 - \alpha)$ quantile of this distribution with $\alpha = 0.01$ for non-CAT item types.

evaluate the issue when enough risky sessions were surfaced within a day. The system relies on two alert thresholds. A session-level threshold α marks any session whose NLL exceeds the α -quantile of the training NLL distribution as “risky.” A day-level alert is triggered when the proportion of risky sessions for an item type i on a day d exceeds $K \times \alpha$, where K is the alert threshold. Both K and α are calibrated from a tuning dataset that includes sessions with known irregularity labels. Specifically, we evaluate a grid of candidate (α, K) pairs against these labels and select the pair that indicate the best tradeoff between recall and the false positive alert rate, subject to a cap on the number of sessions surfaced per false alert (to keep the review burden manageable). Figure 6 shows the training risk-score distribution and the resulting session-level threshold for Interactive Speaking. This same fixed threshold is then applied to the empirical joint densities of this item type from each single day of incoming sessions under evaluation, as illustrated in Figure 7.

In the deployed system, K is not fixed but is recomputed on each scoring day to adjust for seasonal population shifts in test-takers. Concretely, for item type i on scoring day d , let $p_{i,d}$ denote the proportion of that day’s sessions flagged at threshold α_i , and let the trailing baseline be the mean flag rate over the preceding W days:

$$\bar{r}_{i,d} = \frac{1}{W} \sum_{d'=d-W}^{d-1} p_{i,d'}.$$

The alert multiplier applied on that day is

$$K_{i,d} = \max\left(K_i, \left\lceil \frac{M \cdot \bar{r}_{i,d}}{\alpha_i} \right\rceil\right),$$

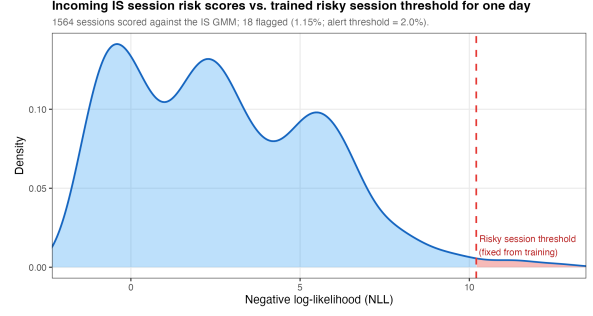


Figure 7: Applying trained risky session thresholds to evaluate the Interactive Speaking response for incoming sessions of a single day. Sessions whose NLL values are to the right of the threshold (red) are surfaced for review.

where $M \geq 1$ is a sensitivity multiplier. Since an alert is triggered when $p_{i,d} > K_{i,d} \times \alpha_i$, and $K_{i,d} \times \alpha_i \geq M \cdot \bar{r}_{i,d}$, the rule requires the day’s proportion of surfaced sessions to exceed the recent baseline by at least a factor of M , with the tuned threshold K_i acting as a floor. M is set through a grid search against the number of false positive alerts in historical tuning data. A gradual population shift raises the acceptable surfaced session rates and the alert threshold together, while an abrupt delivery irregularity raises $p_{i,d}$ faster than the trailing window average and can still trigger an alert. This rolling- K mechanism reduces false alerts due to seasonal population shifts in test takers (Liao et al., 2022) without requiring model retraining. An illustration of the moving alert thresholds for the audio-interactive comprehension item type is shown in Figure 8. Another system design is to conduct periodic retraining of the GMMs using more recent session data, which would update the models’ learned distribution of normal test-taking behaviors and the tuned alert thresholds, so that the reference distribution tracks intended test updates and gradual change in the test-taking population.

Surfaced risky sessions are displayed through a web dashboard, where the top 10-15 sessions with the highest NLL risk scores will be reviewed if an alert takes place. The performance of our alert system is evaluated through a historical dataset, which we discuss in the next section.

3.3 Performance Evaluation

We evaluated AQUA-TT against historical pilot test sessions that had been reviewed and labeled as containing irregularities in item delivery. Table 1 summarized the evaluation results. For two

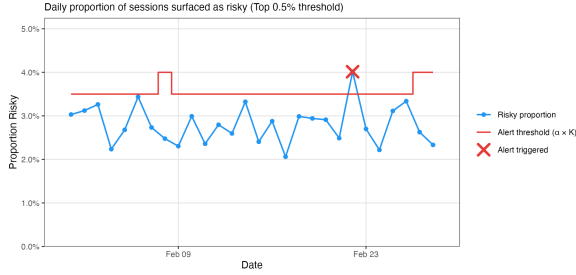


Figure 8: Daily proportion of audio-interactive comprehension sessions surfaced as risky (blue) against the rolling K alert threshold (red step function) in February 2026. On 22 February, 4.0% of AIC sessions were surfaced as risky, exceeding a 3.5% threshold and indicated an alert for human review.

Incident	Item pool	Triggered days (TP:FP)	Days to detection	Alert-day precision
Incident 1	CAT	4 (4:0)	8	1.00
Incident 2	CAT	2 (2:0)	3	1.00
Incident 3	non-CAT	3 (1:2)	7	0.33

Table 1: AQUA-TT detection performance across three labeled historical incidents. Triggered days (TP:FP) = number of days where the system triggered a true or false positive alert within the incident window. Days to detection = days from incident onset to the first true-positive alert. Alert-day precision = fraction of triggered days confirmed as true positives.

incidents affecting CAT item types, the system correctly alerted within 3–8 days of onset, such that each alert day, the sessions truly affected by the incident were among the top sessions the system surfaced for review. A third incident affected a non-CAT item type for under 1% of total daily sessions. The system triggered its first true positive alert within 7 days, though with a lower precision: only 1 of 3 triggered alerts during the incident window contained confirmed affected sessions.

An out-of-sample evaluation was conducted over a 91-day window between April to June 2026, which falls outside the system training and tuning period and contains no documented irregularities. During this period the system produced 10 alerts on production item types (writing and speaking). Seven of these happened within 3 days and were attributed to increased low-effort responses from a new pilot study cohort rather than any test delivery anomaly. The remaining three were isolated false positive alerts. Taken together, these results suggest that the item-type-specific GMM method can validly detect delivery irregularities without incurring inflated false positive alerts. The AQUA-TT

system achieves automated session-level monitoring to support human review of anomalies and ensures the fairness and validity of the DET’s scores. AQUA-TT operationalizes Goals 3.3, 4.1, and 4.3 of the DET RAI Standards, which calls for psychometric procedures to “track anomalies in test-taker performance” and human expert support through “system UX, alert and reporting functions, and rubrics” (Burstein, 2026). The application of our item-specific anomaly detection models and automated alert system for human reviews instantiate these guidelines.

4 Conclusions

We presented two infrastructure systems to demonstrate how the DET follows RAI guidelines as a digital-first assessment. The Item Factory combines automated item generation with staged expert review, and AQUA-TT pairs unsupervised item-level anomaly detection with an alert system to prompt human review. Both systems place human judgment at the end of automated processes in the assessment pipeline. Together they achieve automated and continuous quality assurance at scale in item generation and individual test administration, which are key areas of AI-driven assessment programs (von Davier et al., 2026).

More work can be done to further extend these systems. These include developing alert thresholds and systems for AQUA-TT that are robust to meaningful test changes or population shifts and evaluation methods for Item Factory when ground-truth labels are scarce. In addition, the performance outcomes of both systems are still constrained by reviewer resources. The Item Factory is constrained by how many generated items experts can evaluate, and AQUA-TT’s alert thresholds assume that only a limited number of sessions can be reviewed per alert. As these systems handle an increasing level of operational complexities (a growing item bank and diversifying testing population), it remains to be explored how to further improve the systems’ capacity within the operational constraints and RAI guidelines.

References

- AERA, APA, and NCME. 2014. *Standards for Educational and Psychological Testing*. American Educational Research Association, Washington, DC.
- Yigal Attali, Andrew Runge, Geoffrey T. LaFlair, Kevin P. Yancey, Sarah Goodwin, Yena Park, and

- Alina A. von Davier. 2022. [The interactive reading task: Transformer-based automatic item generation](#). *Frontiers in Artificial Intelligence*, 5:903077.
- William Belzak, Chenhao Niu, and Angel Ortmann Lee. 2025. When machines mislead: Human review of erroneous AI cheating signals. In *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con), Volume 2: Works in Progress*, pages 90–97. National Council on Measurement in Education.
- Christophe Biernacki, Gilles Celeux, and Gérard Govaert. 2000. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(7):719–725.
- Jill Burstein. 2026. [The Duolingo English Test responsible AI standards](#). Technical Report DRR-26-02, Duolingo, Inc.
- Jill Burstein, Geoffrey T. LaFlair, Kathleen Yancey, Alina A. von Davier, and Rotem Dotan. 2025. Responsible AI for test equity and quality: The Duolingo English Test as a case study. In Earl M. Tucker, Eleanor Armour-Thomas, and Edmund W. Gordon, editors, *Handbook for Assessment in the Service of Learning, Volume I: Foundations for Assessment in the Service of Learning*. University of Massachusetts Amherst Library Press.
- Varun Chandola, Arindam Banerjee, and Vipin Kumar. 2009. [Anomaly detection: A survey](#). *ACM Computing Surveys*, 41(3):1–58.
- Jacqueline Church, Yena Park, and Jill Burstein. 2025. [Guidelines for fair test content: The Duolingo English Test example](#). Technical report, Duolingo, Inc.
- Casey Fiesler and Natalie Garrett. 2020. [Ethical tech starts with addressing ethical debt](#). *Wired*.
- Thomas Langenfeld, Jill Burstein, and Alina A. von Davier. 2022. [Digital-first learning and assessment systems for the 21st century](#). *Frontiers in Education*, 7:857604.
- Manqian Liao, Yigal Attali, J. R. Lockwood, and Alina A. von Davier. 2022. [Maintaining and monitoring quality of a continuously administered digital assessment](#). *Frontiers in Education*, 7:857496.
- Rob R. Meijer and Edith M. L. A. van Krimpen-Stoop. 2009. [Detecting person misfit in adaptive testing](#). In Wim J. van der Linden and Cees A. W. Glas, editors, *Elements of Adaptive Testing*, pages 315–329. Springer.
- Ben Naismith, Ramsey Cardwell, Geoffrey T. LaFlair, Steven Nydick, and Maria Kostromitina. 2026. [Duolingo English Test: Technical manual](#). Technical report, Duolingo, Inc.
- Chenhao Niu, Yong-Siang Shih, Manqian Liao, Ruidong Liu, and Angel Ortmann Lee. 2025. Keystroke analysis in digital test security: AI approaches for copy-typing detection and cheating ring identification. In *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con), Volume 2: Works in Progress*, pages 107–116. National Council on Measurement in Education.
- Luca Scrucca, Michael Fop, T. Brendan Murphy, and Adrian E. Raftery. 2016. mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, 8(1):289–317.
- Jorge N. Tendeiro and Rob R. Meijer. 2012. A CUSUM to detect person misfit: A discussion and some alternatives for existing procedures. *Applied Psychological Measurement*, 36(5):420–442.
- Edith M. L. A. van Krimpen-Stoop and Rob R. Meijer. 2002. [Detection of person misfit in computerized adaptive tests with polytomous items](#). *Applied Psychological Measurement*, 26(2):164–180.
- Alina A. von Davier, Andrew Runge, Yena Park, Yigal Attali, Jacqueline Church, and Geoffrey T. LaFlair. 2024. [The item factory: Intelligent automation in support of test development at scale](#). In Hong Jiao and Robert W. Lissitz, editors, *Machine Learning, Natural Language Processing, and Psychometrics*. Emerald Publishing Limited.
- Alina A. von Davier, Xiaowan Zhang, Yigal Attali, Yena Park, Jacqueline Church, Andrew Runge, Geoffrey T. LaFlair, and Alexander Tsigler. 2026. Analytics for quality assurance for item pools (AQuAP): Monitoring and maintaining item bank health in AI-driven assessment systems. *arXiv preprint arXiv:2606.18536*.

Appendix A. Case Study 1 Tables

Batch	Overall		Jargon		Topic		Accessibility		Logic		Stereotype	
	acad.	narr.	acad.	narr.	acad.	narr.	acad.	narr.	acad.	narr.	acad.	narr.
Batch 1	23.26%	21.62%	16.28%	0.00%	9.30%	8.11%	2.33%	5.41%	0.00%	10.81%	2.33%	2.70%
Batch 2	12.83%	15.38%	2.75%	0.00%	2.75%	3.85%	0.92%	0.00%	2.75%	7.69%	0.00%	0.00%

Table A1: Passage Failure Rates by Failure Tag

Passage type	Batch 1 edit dist.	Batch 2 edit dist.	Model-estimated batch difference, p
Academic	13.96%	6.71%	−0.04, .002
Narrative	31.37%	19.54%	−0.04, .285
Overall	22.10%	9.12%	−0.08, < .001

Table A2: Mean Normalized Character Edit Distance Across Batches: Mixed-Effects Model Results

Note. Normalized character edit distance is character-level edit distance divided by the original text length. Percentages represent editing magnitude.

Batch	# Presented for human review		# Selected for operational testing		Mean normalized character edit distance		
	Alt.	Direct	Alt.	Direct	Alt.	Direct	Model-estimated method difference, p
Batch 3	578	80	389 (67.30%)	61 (76.25%)	17.33%	12.50%	−0.03, .265
Batch 4	284	681	182 (64.08%)	520 (76.36%)	22.28%	11.24%	−0.07, < .001

Table A3: Distractor Review and Editing Results

Note. Values represent mean editing magnitude for distractors. The percentage selected for operational use is reported separately.

Batch	% Test-taker selection			Point biserials		
	Alt.	Direct	Comparison, Δ , p	Alternative	Direct	Comparison, Δ , p
Batch 3	5.80%	9.52%	3.72%, < .001	−0.23	−0.26	−.031, .002
Batch 4	6.65%	8.50%	1.85%, < .001	−0.22	−0.25	−.021, < .001
Overall	6.35%	8.61%	2.26%, < .001	−0.23	−0.25	−.021, < .001

Table A4: Distractor Attractiveness and Discrimination Comparison Results

Appendix B. Integrated Complete-data Likelihood

For a candidate GMM model with G components and covariance structure m fit to n training sessions, we use Integrated Complete-data Likelihood (ICL) to select G and m . ICL adds an entropy penalty to the Bayesian Information Criterion $\text{BIC}(G, m)$:

$$\text{ICL}(G, m) = \text{BIC}(G, m) - 2 \text{EN}(G, m),$$

$$\text{where } \text{EN}(G, m) = - \sum_i \sum_g \hat{z}_{ig} \log \hat{z}_{ig}.$$

Here $\text{BIC}(G, m) = 2 \log L - \nu_{G,m} \log n$, where L is the likelihood and $\nu_{G,m}$ is the number of free parameters. $\text{EN}(G, m)$ is the estimated mean entropy of the classification, where $\hat{z}_{ig}$ is the estimated posterior probability that session i belongs to component g . ICL penalizes ambiguous assignment of sessions to components (Biernacki et al., 2000). We select the (G, m) pair that maximizes ICL. Relative to BIC, this yields GMMs with fewer components and results in fewer false positive alerts for the system.