When Machines Mislead: Human Review of Erroneous AI Cheating Signals

William Belzak **Chenhao Niu** **Angel Ortmann Lee**
Duolingo, Inc.
{wbelzak, chenhao, angel.ortmannlee}@duolingo.com

Abstract

Artificial intelligence (AI) systems are increasingly used to monitor high-stakes online exams, but false positives raise concerns about fairness and validity. To study how human reviewers handle erroneous AI alerts, we intentionally faked "copy-typing" signals and embedded them into authentic exam sessions without proctors' awareness. In two experiments, proctors evaluated these fake signals as part of their normal review process. Study 1 established baseline rejection rates, while Study 2 tested revised guidelines emphasizing corroborating evidence of misconduct. Proctors rejected many fake signals (50–71%), though a notable percentage were still accepted. Rejection rates varied somewhat across test-taker nationalities, and the revised guidelines were associated with more consistent decisions across groups. Guideline updates significantly increased rejections of fake signals but also modestly increased rejections of genuine ones, reflecting a tradeoff between reducing false positives and avoiding false negatives. These findings demonstrate the importance of clear guidance and structured oversight in supporting effective human–AI collaboration in exam security.

1 Introduction

For high-stakes exams, test security involves the deterrence, prevention, and detection of cheating and other forms of misconduct that may artificially inflate a test taker's performance beyond their true proficiency. Breaches in security can undermine the validity of test results and carry serious consequences for examinees and other stakeholders, such as deportation or imprisonment (Main and Watson, 2022; McCray, 2019). As such, test security is essential to ensure that stakeholders can accurately interpret and use test scores (AERA et al., 2014).

In recent years, high-stakes exams have increasingly moved online (Weiner and Hurtz, 2017), reducing costs and broadening access for test takers. This shift, however, introduces new security challenges for providers, from controlling the digital test environment to countering sophisticated technological threats. At the same time, AI creates new opportunities for exam security by providing consistent monitoring at scale, enabling the detection of potential misconduct across large numbers of test sessions more efficiently than human proctors alone.

AI is used in many ways to secure high-stakes remote exams (Dawson, 2020; Zenisky and Sireci, 2021), from verifying identities through facial recognition, keystroke tracking, and voice analysis (Nigam et al., 2021) to proctoring tasks such as flagging when a test taker looks away from the screen (Shih et al., 2024) or when another person is detected in the room. It can also monitor for unauthorized devices, unusual movements, or suspicious sounds, and is increasingly applied to detect plagiarism (Liao et al., 2023), AI-generated answers (Niu et al., 2024), or copy-typing behaviors (Niu et al., 2025).

While AI tools can detect many forms of misconduct, they are not perfect: studies show they can misinterpret benign behaviors as misconduct, raising concerns about accuracy (Tweissi et al., 2022), fairness (Yoder-Himes et al., 2022), and student privacy (Balash et al., 2021). Incorporating human oversight, such as having trained proctors review AI-generated signals, can help reduce false positives (Tweissi et al., 2022) and ensure that decisions about potential misconduct are made with appropriate context. Consistent with Responsible AI standards (Burstein, 2025), we highlight the need to balance technical reliability with human oversight.

2 Background

The rapid expansion of remote testing has made security protocols a central concern. The Duolingo English Test (DET) provides a valuable case study

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 2: Works in Progress, pages 90–97

because it integrates AI-based monitoring with human review, offering a real-world setting to evaluate how human decision-makers interact with AI signals in a high-stakes assessment.

2.1 Application

The DET is a remotely administered, high-stakes assessment of English proficiency (Naismith et al., 2025). To protect score integrity, DET employs multiple security measures (Belzak et al., 2025a), including a lock-down digital environment, multi-layered ID verification, an adaptive test design, and standardized administration procedures. Human proctors review test-taker behavior through audio-visual recordings and validate AI-generated signals, ensuring that DET scores remain both reliable and credible. Ultimately, proctors retain final authority in determining whether testing rules were violated or misconduct occurred.

2.2 Copy-Typing Detection

Within the DET security framework, one important safeguard is the detection of potential copy-typing behavior. The DET employs an AI model that analyzes keystroke dynamics, treating a test taker's typing as a sequence of keystroke events and extracting features such as key press durations and intervals between keystrokes. The model architecture captures both local rhythmic patterns and global characteristics across an entire response, enabling it to distinguish between organic composition and transcription from an external source. With the decision threshold selected for this experiment, the estimated false positive rate among test takers who were not copy-typing is about 1%, which represents the upper bound of false positives in the absence of human validation. A full description of the model architecture, feature engineering, and performance is provided in Niu et al. (2025). Although the model has demonstrated strong performance overall, this study did not directly evaluate its construct validity or confidence intervals; readers are referred to prior work for detailed validation (Niu et al., 2025).

When the model flags a test session, the alert is not acted upon automatically. Instead, it is routed to a trained human proctor for review. Proctors are instructed to treat the AI signal as a preliminary alert rather than proof of misconduct. Their primary responsibility is to independently review the audio-visual recording of the session and look for corroborating evidence of cheating as outlined in the proctoring guidelines (see Table 1). Proctors may accept the signal if independent evidence is found, or reject it if no such evidence exists. This ensures a human-in-the-loop validation process.

Even with strong model performance, a small false positive rate can have serious implications in high-stakes testing if unverified alerts are upheld. It is therefore essential to assess whether human proctors can reliably identify and reject false positives, a key safeguard for fairness and test taker protection.

2.3 Research Questions

In this paper, we have three research questions:

1. What percentage of fake copy-typing signals are correctly rejected by proctors?

2. Do rejection rates for fake copy-typing signals differ across test-taker nationalities?

3. Does revising the proctoring guidelines change the likelihood that proctors reject fake copy-typing signals?

We now report on an experiment that aims to answer these questions.

3 Experiment

For this experiment, proctors were asked to accept or reject AI-generated signals indicating potential copy-typing behavior. All signals were intentionally faked, meaning that no test takers were actually flagged for misconduct. These test takers had already received certified DET scores following the standard security review process, ensuring that their results were unaffected.

Fake copy-typing signals were interspersed with real sessions and presented to proctors without their knowledge that any alerts had been faked. Because DET proctoring occurs only after a test is completed, this design was feasible: proctors approached these sessions as part of their normal review process, unaware that the experiment was underway. This allowed us to capture authentic decision-making behaviors under realistic operational conditions.

We ran this experiment twice, first to establish a baseline of rejection rates (Study 1), and second to evaluate how revised proctoring guidelines might have changed those rates (Study 2). Study 1 was conducted from January 30 to February 13, 2025, using the original guidelines. On March 28, 2025,

the guidelines were revised to remove "irregular typing patterns" as a criterion, add "presence of an external resource," and to instruct proctors not to apply a copy-typing flag unless suspicious behaviors were observed in the video evidence. These changes were intended to reinforce that proctors should only uphold a copy-typing flag when independent evidence of copy-typing behaviors was present. Study 2 was conducted from July 15 to July 31, 2025, with proctors applying the revised guidelines. Both versions of the guidelines are presented in Table 1.

3.1 Data

For each study, we randomly selected $N = 170$ test sessions that met three conditions: no copy-typing signal was triggered, no misconduct was identified by proctors (all test takers had received certified scores), and the sessions had not been escalated to secondary review for borderline or complex cases. Different sessions were sampled for the two studies because the test content had changed over time and, importantly, to avoid alerting proctors that they might be reviewing the same sessions twice, which could have undermined the realism of the task. Despite being drawn from different time periods, the two sets of sessions showed highly similar distributions of copy-typing detection logits (Figure 1), and the mean values did not differ significantly ($p = 0.433$).

To examine group differences, our analysis focused on three nationality groups—Western (American, Canadian, and French), Chinese, and Indian—as they represented the largest test-taker populations in both our dataset (see Table 2) and the DET (Michalowski et al., 2024). Nationality was selected as a key variable for evaluating group differences in copy-typing decisions, as prior research shows that proctoring outcomes are especially sensitive to this factor. For example, Belzak et al. (2025b) found that both proctor and test-taker nationality influenced the likelihood of being flagged for rule violations, whereas other demographic characteristics such as gender and age did not. Table 3 presents the broader distribution of proctor nationalities aggregated by continent. These data reveal that the majority of proctors are based in the Americas (46%) and Europe (24%), a distribution that contrasts with the larger populations of Chinese and Indian test takers.

In addition to the fake signals, we also collected operational sessions that had been flagged with

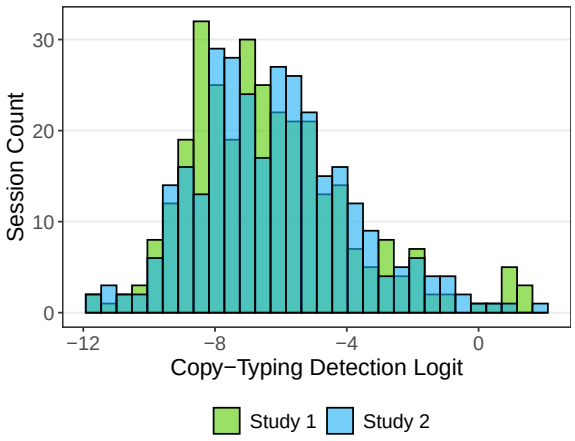

Figure 1: Distribution of copy-typing detection model predictions for sessions used in two studies, with the average values and standard deviations being -6.20 ± 2.36 for Study 1 and -6.36 ± 2.46 for Study 2.

genuine copy-typing signals during the two periods before and after the guideline revision. These data allowed us to examine whether changes in proctoring instructions influenced how proctors handled authentic AI alerts, providing a real-world complement to the experimental results based on fake signals.

3.2 Methods

We first estimate the probability of proctors rejecting fake copy-typing signals by fitting a logistic mixed-effects model (Raudenbush and Bryk, 2002) to the combined data from both studies:

$$\text{logit}\big(Pr(S_{ij} = 0)\big) = \beta_0 + \beta_1 X_i + u_j, \quad (1)$$

where $Pr(S_{ij} = 0)$ is the probability that proctor j rejects signal i, $X_i = 0$ under the original guidelines and $X_i = 1$ under the revised guidelines, and $u_j \sim N(0, \tau^2)$ is a random effect for proctor j, which accounts for non-independence because each proctor evaluated multiple signals. In this model, β_0 represents the baseline log-odds of rejection under the original guidelines, while β_1 captures the log-odds change after the revision.

Next, we examine nationality effects in two stages. In the first stage, we fit within-study models to test for differences among Chinese, Indian, and Western test takers:

$$\text{logit}\big(Pr(S_{ij} = 0)\big) = \beta_0 + \beta_1 C_i + \beta_2 I_i + u_j, \quad (2)$$

where $C_i = 1$ for Chinese test takers and 0 otherwise, $I_i = 1$ for Indian test takers and 0 otherwise, with Western test takers (American, Canadian, French) as the reference group. Here, β_0

Original Guidelines	Revised Guidelines
Review corresponding video segments for suspicious behaviors that indicate copy-typing, including: • Irregular Typing Patterns • Unusual Body Movements • Irregular Eye Movements	Review corresponding video segments for suspicious behaviors that indicate copy-typing, including: • ~~Irregular Typing Patterns~~ • Unusual Body Movements • Irregular Eye Movements • *Presence of an external resource* ***Do not apply** this signal flag if no suspicious behaviors are observed.*

Table 1: Proctoring guidelines for reviewing copy-typing signals. Revisions are marked as ~~deleted~~ or *added*.

Nationality Group	Study 1	Study 2
Western*	25	30
Chinese	28	17
Indian	15	19

Table 2: Number of test takers by nationality group in Study 1 and Study 2. The Western nationality group includes American, Canadian, and French test takers.

Continent	Percentage
Americas	46%
Europe	24%
Asia	13%
Africa	10%
Oceania	6%

Table 3: Percentage of proctor nationalities aggregated by continent.

gives the baseline log-odds of rejection for Western test takers, while β_1 and β_2 capture contrasts for Chinese and Indian test takers, respectively. In the second stage, we use the same model specification as Eq. 1, but fit it separately within each nationality group to assess between-study differences. This allows us to test whether rejection rates changed significantly from Study 1 (original guidelines) to Study 2 (revised guidelines) within each nationality.

Finally, we estimate the probability of rejecting *genuine* copy-typing signals observed during operational proctoring. The model includes the guideline condition (original vs. revised) as a predictor, specified in the same way as Eq. 1. Unlike Eq. 1, however, we use a logistic regression model rather than a mixed-effects model, since each genuine signal was reviewed by only one proctor.

For all models, model-implied probabilities and percentages are obtained by applying the inverse-logit transformation to the estimated log-odds coefficients.

4 Results

Figure 2 shows the model-implied percentages of proctors rejecting fake copy-typing signals under the original and revised guidelines. Results from the logistic mixed-effects model indicate that rejection rates were significantly higher after the revision, with an estimated effect of $\hat{\beta}_1 = 0.880$ ($p = .001$) on the log-odds scale.

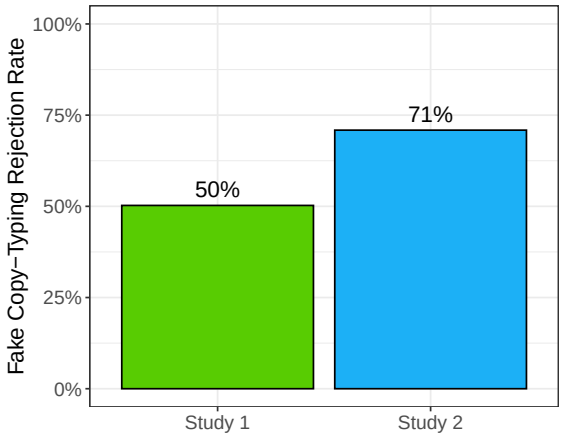

Figure 2: Model-implied percentages of rejecting fake copy-typing signals before (Study 1) and after (Study 2) the revision of proctoring guidelines.

Figure 3 shows the model-implied percentages of proctors rejecting fake copy-typing signals by test-taker nationality and study. The within-study logistic mixed-effects models revealed systematic nationality differences. In Study 1, Chinese test takers were significantly less likely than Western test takers to have fake signals rejected ($\hat{\beta}_1 = -2.230$, $p < .001$). Indian test takers also showed lower rejection rates than Western test takers, though this difference was marginally statistically significant ($\hat{\beta}_2 = -1.159$, $p = .088$). In Study 2, Chinese test takers again exhibited lower rejection rates than Western test takers, but the difference was not significant ($\hat{\beta}_1 = -0.996$, $p = .187$). By contrast, Indian test takers were significantly less likely than Western test takers to have fake signals rejected ($\hat{\beta}_2 = -1.553$, $p = .029$).

The between-study logistic mixed-effects models also revealed systematic differences by nationality. Chinese test takers showed significantly higher rejection rates in Study 2 compared to Study 1 ($p = .001$). Rejection rates for Western and Indian test takers also increased across studies, but these effects did not reach statistical significance ($p = .104$ and $p = .303$, respectively).

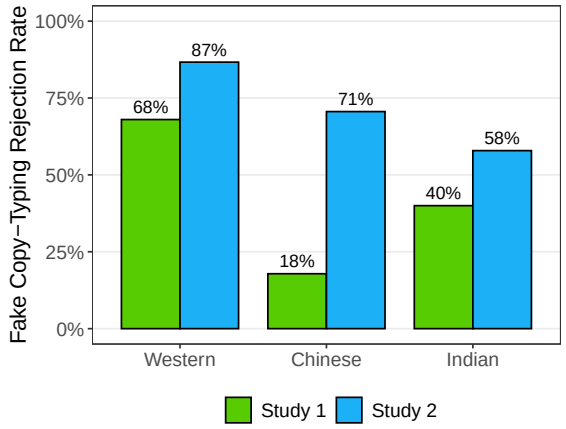

Figure 3: Model-implied percentages of rejecting fake copy-typing signals by test-taker nationality and study.

Figure 4 shows the model-implied percentages of proctors rejecting both fake and genuine copy-typing signals under the original and revised guidelines. Rejection rates for fake signals were substantially higher (50–71%) than for genuine signals (9–13%). The logistic model also indicated a small but statistically significant increase in the rejection of genuine signals after the revision ($p = .003$). We discuss the implications of these findings in the next section.

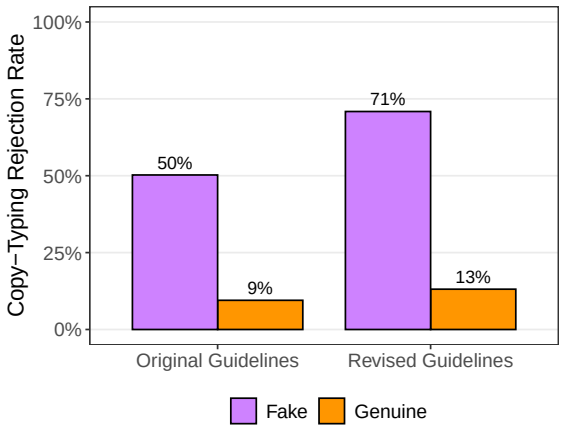

Figure 4: Model-implied percentages of rejecting fake and real copy-typing signals under the original and revised guidelines.

5 Discussion

This experiment examined how trained human proctors interact with AI-generated copy-typing signals in a remotely administered, high-stakes English language assessment. By intentionally fabricating signals and asking proctors to accept or reject them under specific guidelines, we evaluated three questions: (1) whether proctors could correctly reject fake AI signals, (2) whether rejection rates varied by test-taker nationality, and (3) whether revised guidelines changed proctoring decisions.

5.1 Research Question 1: Percentage of Fake Signal Rejections

The first research question asked: *What percentage of fake copy-typing signals are correctly rejected by proctors?* Across both studies, proctors rejected a substantial percentage of fake signals (50–71%). This suggests that proctors can often identify when an AI alert is not supported by independent evidence of misconduct. However, the fact that roughly one in three fake alerts was accepted underscores the risks of overreliance on AI signals in high-stakes contexts (Skitka et al., 1999; Poursabzi-Sangdeh et al., 2021). These findings highlight the value of human review for maintaining fairness and accuracy, while also pointing to the need for additional safeguards to minimize the consequences of false positives (Bansal et al., 2021). As AI detection models improve, false positive rates, and the role of human reviewers in rejecting them, may shift. More research will be critical to ensure systems remain both reliable and fair.

5.2 Research Question 2: Nationality Differences in Rejection Rates

The second research question asked: *Do rejection rates for faake copy-typing signals differ across test-taker nationalities?* The within-study analyses revealed systematic differences: in Study 1, Chinese test takers were significantly less likely to have fake signals rejected than Western test takers, and Indian test takers showed a similar trend. In Study 2, rejection rates for Chinese test takers improved and were no longer significantly different from Western test takers, while Indian test takers were significantly less likely to have fake signals rejected. The between-study analyses confirmed that rejection rates increased significantly for Chinese and Western test takers in Study 2, but not for Indian test takers. These findings suggest that proctor decision-making can vary by nationality group, perhaps due to a structural mismatch between the distribution of test takers and proctors (Belzak et al., 2025b), and that revised guidelines may reduce some differences while leaving others unaddressed. Strategies such as targeted proctor training, bias monitoring dashboards, and regular fairness audits could help ensure that future revisions to proctoring guidelines improve accuracy while also addressing inequities across groups.

5.3 Research Question 3: Effect of Revised Guidelines

The third research question asked: *Does revising the proctoring guidelines change the likelihood that proctors reject fake copy-typing signals?* The evidence indicates that they do. After the guidelines were updated to emphasize the need for corroborating evidence of misconduct, rejection rates of fake signals increased significantly. This suggests that proctoring practices are sensitive to instructional framing and that targeted revisions can improve decision quality (Association of Test Publishers and National College Testing Association, 2024; Buçinca et al., 2021). However, the revised guidelines were also associated with a small but statistically significant increase in the rejection of genuine signals, suggesting that proctors became more cautious about accepting AI alerts but also more likely to dismiss valid cases (Almog et al., 2024). This tradeoff between reducing false positives and increasing false negatives highlights the complexity of calibrating human–AI collaboration in high-stakes testing. More research is needed to refine this balance and identify guidelines that reduce risks without undermining security.

5.4 Implications

Taken together, the findings underscore both the value and limitations of human-in-the-loop AI systems in exam security. Proctors are capable of rejecting false positive copy-typing signals, but not always uniformly across nationalities, and their decisions are shaped by the guidance they receive. Ongoing training, carefully designed guidelines, and continuous monitoring of decision patterns are therefore essential to ensure fairness and validity (Burstein et al., 2025).

5.5 Limitations and Future Work

Several limitations should be considered when interpreting these findings. First, the two studies were conducted on different sets of test sessions and several months apart. This was necessary because the test's visual design had changed, and reusing the same sessions could have signaled to proctors that they were artificial. However, this design also means that unobserved differences in session characteristics or other contextual changes over time may have contributed to the observed effects, making it difficult to attribute differences solely to the revised guidelines.

Second, for genuine copy-typing signals observed during operational proctoring, we cannot determine whether higher rejection rates reflect proctors dismissing false positives or overlooking true positives. As such, estimates of genuine-signal rejection rates should be interpreted with caution. Establishing verified ground truth through simulated or confirmed cases of misconduct would strengthen future studies.

Third, the scope of this experiment was limited in terms of sample size, signal type, and use of fake copy-typing signals. The relatively small samples constrained analyses of nationality differences, and fake signals—while useful for preserving realism—may not capture the full complexity of genuine AI alerts. Moreover, we focused on copy-typing signals only; other alerts, such as those for unusual movements, unauthorized devices, or suspicious sounds, may pose different challenges for human validation. Future work should expand to larger, more diverse datasets and a broader range of signal types to better understand the dynamics of human–AI collaboration in exam security.

Addressing these limitations will be essential

for improving both the accuracy and fairness of AI-assisted proctoring. Larger datasets, verified ground truth, and broader signal coverage will help test providers calibrate human–AI decision-making and safeguard the integrity of high-stakes assessments.

6 Conclusion

This study examined how trained human proctors interact with AI-generated copy-typing signals in a high-stakes, remotely administered English language assessment. To ensure operational outcomes were unaffected, fake signals were embedded only into completed sessions where test takers had already received certified scores. This design allowed us to evaluate proctor decision accuracy in rejecting fake AI signals, explore differences across nationality groups, and assess the impact of revised guidelines without altering test results.

Proctors generally identified and rejected fake copy-typing signals, but acceptance of some signals highlights the risks of overreliance on AI. Rejection rates varied by nationality, with differences reduced but not eliminated under revised guidelines. The guidelines also increased rejections of fake signals while slightly raising rejections of genuine ones, underscoring the tradeoff between false positives and false negatives. More research is needed to examine these dynamics in larger datasets, across different AI signals, and in varied testing contexts.

Overall, the findings illustrate both the promise and limits of human-in-the-loop AI for exam security. Clearer guidelines, regular training, and monitoring are essential to support fairness and validity. Practically, testing organizations can refine proctor training, track nationality-related outcomes, and calibrate AI–human collaboration to balance accuracy and fairness. Because the reliability of human–AI systems depends not only on technical performance but also on governance, transparent processes, and oversight, exam security frameworks should be aligned with Responsible AI standards (Burstein et al., 2025). More broadly, stakeholder trust in high-stakes assessments also rests on adherence to the principles of fairness and validity articulated in the *Standards for Educational and Psychological Testing* (AERA et al., 2014).

Acknowledgments

Generative AI was used to review and refine text originally drafted by the authors, improving readability and flow. All content was verified for accuracy. We thank Alina von Davier and Jill Burstein for their thoughtful feedback on the draft.

References

AERA, APA, and NCME. 2014. *Standards for Educational and Psychological Testing*. American Educational Research Association, Washington, DC.

David Almog, Romain Gauriot, Lionel Page, and Daniel Martin. 2024. AI oversight and human mistakes: evidence from centre court. In *Proceedings of the 25th ACM Conference on Economics and Computation*, pages 103–105.

Association of Test Publishers and National College Testing Association. 2024. Assessment industry standards and best practices for the online observation of tests. https://www.testpublishers.org/assets/Online%20Observation%20of%20Tests%20Standards%20for%20Public%20Comment%202024.3.23.1207%20numbered.pdf. Association of Test Publishers & National College Testing Association.

David G Balash, Dongkun Kim, Darika Shaibekova, Rahel A Fainchtein, Micah Sherr, and Adam J Aviv. 2021. Examining the examiners: Students' privacy and security perceptions of online proctoring services. In *Seventeenth symposium on usable privacy and security (SOUPS 2021)*, pages 633–652.

Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of AI explanations on complementary team performance. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–16.

William Belzak, Basim Baig, Ramsey Cardwell, Rose Hastings, Andre Horie, Geoff LaFlair, Manqian Liao, Chenhao Niu, and Yong-Siang Shih. 2025a. Duolingo english test: Security and score integrity. Duolingo Research Report DRR-25-04, Duolingo English Test.

William Belzak, Jill Burstein, and Alina A. von Davier. 2025b. Evaluating fairness in AI-assisted remote proctoring. In *Proceedings of the Innovation and Responsibility in AI-Supported Education Workshop*, volume 273 of *Proceedings of Machine Learning Research*, pages 125–132.

Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-computer Interaction*, 5(CSCW1):1–21.

Jill Burstein. 2025. Duolingo english test: Responsible ai standards. Duolingo Research Report DRR-25-05, Duolingo.

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.

Phillip Dawson. 2020. *Defending assessment security in a digital world: Preventing e-cheating and supporting academic integrity in higher education*. Routledge.

Manqian Liao, Sinon Tan, and Baig Basim. 2023. Plagiarism detection using human-in-the-loop AI. *Paper presented at the annual meeting of the National Council on Measurement in Education*.

Ed Main and Richard Watson. 2022. The english test that ruined thousands of lives. https://www.bbc.com/news/uk-60264106. Accessed: 26 August 2025.

Vanessa McCray. 2019. Judge in APS cheating trial to remain on case as six seek retrial. *The Atlanta Journal-Constitution*. Accessed via The Atlanta Journal-Constitution.

Allison Michalowski, Ramsey Cardwell, Steven Nydick, and Ben Naismith. 2024. Duolingo english test: Demographic and score properties of test takers. Technical report, Duolingo Research Report). Duolingo. https://go. duolingo. com/demographic-score.

Ben Naismith, Ramsey Cardwell, Geoffrey T. LaFlair, Steven Nydick, and Masha Kostromitina. 2025. *Duolingo English Test: Technical Manual*. Duolingo, Inc. Last updated July 2025.

Aditya Nigam, Rhitvik Pasricha, Tarishi Singh, and Prathamesh Churi. 2021. A systematic review on AI-based proctoring systems: Past, present and future. *Education and Information Technologies*, 26(5):6421–6445.

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. *Proceedings of Artificial Intelligence in Measurement and Education Conference*.

Chenhao Niu, Kevin P. Yancey, Ruidong Liu, Mirza Basim Baig, André Kenji Horie, and James Sharpnack. 2024. Detecting LLM-assisted cheating on open-ended writing tasks on language proficiency tests. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 940–953, Miami, Florida, US. Association for Computational Linguistics.

Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–52.

Stephen W Raudenbush and Anthony S Bryk. 2002. *Hierarchical linear models: Applications and data analysis methods*, volume 1. sage.

Yong-Siang Shih, Zach Zhao, Chenhao Niu, Bruce Iberg, James Sharpnack, and Mirza Basim Baig. 2024. AI-assisted gaze detection for proctoring online exams.

Linda J Skitka, Kathleen L Mosier, and Mark Burdick. 1999. Does automation bias decision-making? *International Journal of Human-Computer Studies*, 51(5):991–1006.

Adiy Tweissi, Wael Al Etaiwi, and Dalia Al Eisawi. 2022. The accuracy of AI-based automatic proctoring in online exams. *Electronic Journal of e-Learning*, 20(4):419–435.

John A Weiner and Gregory M Hurtz. 2017. A comparative study of online remote proctored versus onsite proctored high-stakes exams. *Journal of Applied Testing Technology*, pages 13–20.

Deborah R Yoder-Himes, Alina Asif, Kaelin Kinney, Tiffany J Brandt, Rhiannon E Cecil, Paul R Himes, Cara Cashon, Rachel MP Hopp, and Edna Ross. 2022. Racial, skin tone, and sex disparities in automated proctoring software. In *Frontiers in Education*, volume 7, page 881449. Frontiers Media SA.

April L. Zenisky and Stephen G. Sireci. 2021. The impact of technology on test security. In Jeffrey K. Smith, editor, *The Oxford Handbook of Educational Measurement*, pages 386–402. Oxford University Press.