Evaluating Fairness in AI-Assisted Remote Proctoring

William Belzak WBELZAK@DUOLINGO.COM
Jill Burstein JILL@DUOLINGO.COM
Alina A. von Davier AVONDAVIER@DUOLINGO.COM
Duolingo, Inc. 5900 Penn Avenue Pittsburgh, PA 15206

Abstract

Remote proctors make decisions about whether test takers have violated testing rules and, as a result, whether to certify test takers' scores. These decisions rely on both AI signals and human evaluation of test-taking behaviors. Given that fairness is a key component of test validity evidence, it is critical that proctors' decisions are unbiased with respect to proctor and test-taker background characteristics (e.g., gender, age, and nationality). In this study, we empirically evaluate whether proctor or test-taker background characteristics affect whether a test taker is flagged for rule violations. Results suggest that proctor and test-taker *nationality* may influence proctoring decisions, whereas *gender* and *age* do not. The direction of the influence generally reflects an "in-group, out-group" bias: proctors are less likely to identify rule violations among test takers with similar nationalities as proctors (in-group favoring) and more likely to identify rule violations among test takers of different nationalities (out-group disfavoring). Results also suggest that decisions based on AI signals may be less prone to in-group/out-group bias than decisions based on human evaluation only, although more research is needed to support this finding.

Keywords: remote proctoring, responsible AI, fairness, in-group/out-group bias, education

1. Introduction

Developing digital, high-stakes assessments involves three key frameworks: test design, measurement, and security. For test design, test developers design question types, leverage generative AI to develop new questions at scale, and employ humans to review content, ensuring test items are unoffensive and relevant to the measured construct. For measurement, psychometricians determine test reliability and develop statistical and machine learning models to score test-taker responses and evaluate questions for bias. For security, test security professionals manage test administration and data privacy, and use forensic analyses and AI to detect potential cheating activity. Cheating activity comes in many forms (Cizek, 1999; Dawson, 2020), from plagiarism and the use of large language models (LLMs) to proxy test-taking and deepfake technology. For remote assessments in particular, human proctors monitor test takers through video- and audio-recording software and evaluate test-taking behaviors for suspicious activity. AI signals are also leveraged to detect plagiarism across large swaths of data (Liao et al., 2023), identify test takers who are using LLMs to respond to test questions (Niu et al., 2024), and assist with tracking eye-gaze movements (Shih et al., 2024), among many other things.

Existing empirical research on fairness in assessment has primarily focused on test design and measurement, but less on test security. Although in-person proctoring has been

scrutinized for potential biases (Winke et al., 2013), remote proctoring introduces unique challenges due to the lack of physical presence, reliance on technology, and diverse global testing populations (Isbell et al., 2023). This setting raises questions about whether biases, such as in-group favoritism, out-group discrimination, or AI-prompted biases, manifest in remote proctoring decisions, particularly when proctors and test takers differ in background characteristics. Research on fairness in online proctoring highlights the ethical challenges posed by the technologies used to support proctors (Dawson, 2024), especially around privacy and transparency. For instance, Coghlan et al. (2021) and Nigam et al. (2021) focus on the ethical concepts of academic integrity, fairness, transparency, privacy, and trust as they relate to remote proctoring and AI. Most of these concepts are also prominent in the new field of AI ethics and fairness Mehrabi et al. (2021), and all are relevant to the educational context. Moreover, this research provides ethical considerations that educational institutions will need to carefully review before electing to deploy and govern specific online proctoring technologies.

2. Case Study

In this paper, we use the Duolingo English Test (DET)—a digital, high-stakes English language proficiency test (Cardwell et al., 2024)—as a case study for investigating proctoring fairness. Notably, the DET leverages the frameworks described above, using AI with human oversight to automate test item creation, scoring, and security. The DET is also remotely administered, meaning that it is taken "at-home" or anywhere where the test taker has a quiet space and internet access. This is an alternative approach to test administration compared to the more conventional approach where test takers need to travel to a test center to take a test. Test security of remote assessments is more highly scrutinized by test stakeholders, such as university administrators who use the test for admissions purposes, in part because it is new and less tested.

Given that the DET relies heavily on AI, responsible AI (RAI) standards and practices are essential for mitigating potential biases due to AI. To that end, the DET focuses on four RAI standards that represent ethical principles aligned with the DET's goals (Burstein, 2023). First, the Validity and Reliability standard ensures the test is suitable for its purpose, with Validity focusing on accuracy and construct relevance, and Reliability ensuring consistency in measurement. The Fairness standard promotes access, accommodations, representative test-taker demographics, and the minimization of algorithmic bias. The Privacy and Security standard ensures legal compliance, test-taker privacy, and secure test administration. Lastly, the Accountability and Transparency standard builds trust through proper governance and documentation of AI usage. These standards provide a comprehensive framework for evaluating AI-powered assessments.

The analysis in this paper aligns most closely with the Fairness standard by investigating whether proctoring decisions are influenced by diverse test-taker demographics and whether human evaluations or AI signals differ in their susceptibility to bias.

3. Empirical Analysis

In this study, we evaluate whether proctoring decisions on the DET may be influenced by proctor or test taker background characteristics. Specifically, we evaluate the impact of proctor and test taker gender, age, and nationality on the likelihood of a proctor flagging a test taker for one or more rule violations under operational settings. We also examine whether proctoring decisions based on AI signals (e.g., test taker was detected using a large language model) are more or less prone to bias than decisions based solely on proctors analyzing test-taking behaviors (e.g., test taker was looking away suspiciously).

3.1. Data

We collected proctoring data ranging from 2023-07-01 to 2024-11-01 based on the DET, which included $N > 2$ million proctoring decisions from $N > 300$ proctors. The DET uses a "record-and-review" approach to remote proctoring, where proctors review video and audio recordings of test takers after the test is completed and then determine whether any rule violations occurred.

Eight different proctoring decisions were evaluated for bias. Seven of the eight decisions represent specific rule violations that the proctor identified, two of which were based on AI signals: "Detected copy-typing" and "Detected using LLM". The other rule violations were determined by the proctor through behavioral analysis of the test taker: "Looking away suspiciously", "Using scripted response", "Using prohibited device", "Not completing written section independently", and "Receives external assistance". Lastly, we evaluated whether the proctor decided a test score should be certified or not, denoted as "OK".

Six different background variables were also evaluated for bias, two based on gender (Male vs. Female) and age (continuous) and the remaining based on nationality: United States (US), United Kingdom (UK), Philippines (PH), and India (IN). Proctor background data was self-reported, whereas test-taker background data was collected from official identification documents.

3.2. Model

Based on Belzak et al. (2024)'s approach, we used logistic mixed-effects models to predict the probability that proctor j made decision d about test taker i, denoted $\Pr(d = 1)_{ij}$, based on an interaction between proctor and test-taker background characteristics, denoted P_j and T_i. This is mathematically stated as

$$\Pr(d = 1)_{ij} = \frac{1}{1 + \exp[-(\beta_0 + \beta_1 P_j + \beta_2 T_i + \beta_3 (P_j \times T_i) + u_j]} \tag{1}$$

where β_0 is the expected log-odds decision rate of the proctor and test taker groups coded as 0 (e.g., Male proctors and Male test takers); β_1 is the expected difference in the log-odds decision rate between the proctor group coded as 1 (e.g., Female proctors) and the proctor group coded as 0, controlling for T_i; β_2 is the expected difference in the log-odds decision rate between the test-taker group coded as 1 (e.g., Female test takers) and the test-taker group coded as 0, controlling for P_j; β_3 is the expected difference in the effect of the proctor group coded 1 on the log-odds decision rate as a function of the test-taker group coded 1;

and finally, u_j is the random effect for proctor j, distributed N(0, τ_j) with τ_j denoting the proctor-level variance.

P_j and T_i either represented the same background characteristic for proctors and test takers (e.g., proctor males and test-taker males) or different characteristics across proctors and test takers (e.g., proctors with US nationality and test takers with Indian nationality). In total, 18 different combinations of the background variables were tested, including every proctor-by-test-taker nationality combination (e.g., US proctor and US test-taker, US proctor and Indian test-taker, and so on). With eight proctoring decisions (i.e., model outcomes), we specified 144 (18×8) models.

Based on Equation (1), a proctor's decision may be considered biased if it depends on their background and particularly if the test taker's background strengthens this dependency. In other words, if one or both of the proctor main effect and the proctor-by-test-taker interaction effect (β_1 and β_3) are statistically significant, proctoring bias may be present. Conversely, a proctor's decision may not necessarily be considered biased if it depends only on the test taker's background, controlling for a proctor's background. That is, some groups of test takers may truly violate testing rules more often than other groups of test takers (e.g., test takers with high proficiency may be less likely to cheat than test takers with low proficiency). Thus, for our purposes it does not matter whether the test-taker main effect (β_2) is statistically significant or not.

3.3. Results

In Table 1, we show that 18 out of 144 models (13%) exhibited statistically significant proctor-by-test-taker interaction effects (β_3) and 4 out of 144 models (3%) exhibited statistically significant proctor main effects (β_1), both at the .05 alpha level. When correcting for multiple hypothesis testing via the False Discovery Rate method, only 6 of the 144 interaction effects (4%) and 2 of the 144 proctor main effects remain statistically significant. Of note, proctors' decisions about test takers "looking away suspiciously" represented half of all the statistically significant effects.

We also show in Table 1 the type of bias each effect represents. Most interaction effects show either an in-group (favoring) or out-group (disfavoring) bias, but 3 of the 18 show the out-group being favored. Furthermore, two of the four main effects show a strictness type of bias due to proctor age, where strictness refers to an elevated rate of flagging test takers for rule violations (e.g., older proctors are more likely to identify test takers as using a prohibited device compared to younger proctors). Conversely, the other two main effects show a leniency bias due to proctor US nationality, where leniency refers to a diminished rate of flagging test takers for rule violations (e.g., US proctors are less likely to identify test takers as looking away suspiciously compared to non-US proctors).

4. Discussion

The findings of this study suggest that proctoring decisions in remote assessments may be influenced by patterns that align with in-group and out-group bias (Brewer, 1999; Meeus et al., 2010). Proctors appeared less likely to flag rule violations among test takers with similar backgrounds, and more likely to flag test takers from different backgrounds, particularly when evaluating behaviors through human judgment alone. These tendencies, while subtle,

Table 1: Statistically significant interactions between proctor and test-taker background characteristics

Proctor Background	**Test-Taker Background**	**Decision**	**Effect**	**Estimate (Probability)**	**Type of Bias**
US	PH	OK	Int (β_3)	0.057	*Out-group (favor)*
IN	IN	OK	Int (β_3)	0.045*	In-group (favor)
PH	UK	Looks away suspiciously	Int (β_3)	0.043	Out-group (disfavor)
US	US	OK	Int (β_3)	0.032	In-group (favor)
PH	US	Looks away suspiciously	Int (β_3)	0.029	Out-group (disfavor)
PH	US	Using scripted response	Int (β_3)	0.021	Out-group (disfavor)
PH	IN	Using scripted response	Int (β_3)	0.018*	Out-group (disfavor)
US	IN	Looks away suspiciously	Int (β_3)	0.015*	Out-group (disfavor)
UK	IN	Looks away suspiciously	Int (β_3)	0.010	Out-group (disfavor)
UK	IN	Using prohibited device	Int (β_3)	0.005	Out-group (disfavor)
Age	Age	Using prohibited device	Main (β_1)	0.005*	Strictness
Age	Age	Receiving external assistance	Main (β_1)	0.002*	Strictness
PH	IN	Detected copy-typing	Int (β_3)	0.001	Out-group (disfavor)
Male	Male	Detected using LLM	Int (β_3)	−0.002	In-group (favor)
UK	IN	Using scripted response	Int (β_3)	−0.011*	*Out-group (favor)*
US	IN	OK	Int (β_3)	−0.015*	Out-group (disfavor)
IN	IN	Looks away suspiciously	Int (β_3)	−0.022	In-group (favor)
PH	IN	Looks away suspiciously	Int (β_3)	−0.023*	*Out-group (favor)*
US	UK	Looks away suspiciously	Main (β_1)	−0.027	Leniency
US	IN	Looks away suspiciously	Main (β_1)	−0.028	Leniency
UK	UK	Looks away suspiciously	Int (β_3)	−0.046	In-group (favor)
PH	UK	OK	Int (β_3)	−0.070	Out-group (disfavor)

Note. Asterisked (*) values of "Estimate (Probability)" indicate statistical significance after correcting for False Discovery Rate. Italicized values of "Type of Bias" indicate bias favoring an out-group.

align with prior research on how social and cognitive factors can unconsciously influence decision-making (Bruch and Feinberg, 2017).

For example, Filipino (PH) proctors were slightly more likely to flag US test takers for behaviors like "looking away suspiciously" ($\beta_3 = 0.029$), while UK proctors were less likely to flag UK test takers for the same behavior ($\beta_3 = -0.046$). Indian (IN) proctors also showed a tendency to be more lenient toward Indian test takers when certifying scores ($\beta_3 = 0.045$). These observations may reflect an unconscious preference for individuals perceived as culturally similar or a greater scrutiny of individuals from different backgrounds (Hewstone et al., 2002). However, it is important to note that these patterns may also be influenced by contextual factors, such as differences in behavioral norms or communication styles across cultures (Gudykunst et al., 1996).

The remote nature of proctoring may amplify or dampen these tendencies, as proctors rely on video recordings rather than direct interactions with test takers. Without opportunities to build rapport or interpret behaviors within a broader context, proctors may be more or less likely to draw on implicit assumptions or stereotypes. This underscores the need for careful attention to how proctoring decisions are made, particularly when they involve subjective interpretations of test-taker behavior.

Another notable finding from this study is the difference in how human evaluations and AI-generated signals are associated with potential patterns of in-group or out-group bias. Decisions derived from human evaluations only—such as whether a test taker appeared suspicious by "looking away"—were more likely to reflect these patterns compared to decisions based on AI-generated signals. By contrast, decisions based on AI signals, such as detecting large language model (LLM) usage, showed fewer instances of such patterns. For example, while multiple significant interaction effects were observed for human evaluation decisions, only one case of potential in-group favoring was observed in AI-related decisions. Male proctors were slightly less likely to flag male test takers for LLM use ($\beta_3 = -0.002$). These findings suggest that AI tools, by standardizing decision criteria and focusing on objective data inputs, may reduce the influence of proctor and test-taker background characteristics.

Of course, the neutrality of AI tools cannot be taken for granted (Rau et al., 2009). Their performance and fairness depend heavily on the design of the algorithms and the quality of the training data (Mehrabi et al., 2021). While AI systems appear to mitigate certain biases in this study, careful validation and ongoing monitoring are necessary to ensure these tools provide equitable outcomes for diverse test-taker populations.

5. Limitations and Future Directions

While AI-assisted evaluation shows promise in reducing bias, it may have limitations in interpreting nuanced human behaviors, contextual cues, and cultural variations in test-taking practices (Prabhakaran et al., 2022). AI may struggle with subtle expressions of anxiety, deception, or engagement that human proctors intuitively recognize, although these interpretations may also be prone to more unreliable decision-making. Additionally, the feasibility of standardizing proctor decisions with AI raises ethical concerns—should past decisions dictate future evaluations, and how do we account for evolving social norms? Furthermore, analyzing complex interactions, such as the combined effects of gender and nationality, may provide deeper insights into bias patterns, though dataset constraints may limit such exploration in this study. Finally, the study's findings may not generalize across different testing scenarios (e.g., live remote proctoring), underscoring the need for future research across diverse settings to validate and refine AI-assisted proctoring approaches.

6. Conclusion

The findings from this study suggest that proctoring decisions, particularly those relying on human evaluations, may reflect subtle patterns of in-group favoring or out-group scrutiny. Decisions based on AI-generated signals appear less influenced by these dynamics, highlighting the potential of AI to promote fairness in remote proctoring. However, both human and AI-driven decisions warrant careful attention to ensure fair treatment for all test takers, regardless of their background. Proctors may benefit from training that emphasizes objective criteria for evaluating behaviors, alongside a deeper understanding of cross-cultural norms. For AI systems, regular audits of training data and algorithms are essential to prevent unintended biases from influencing outcomes. Building on these findings and conclusions, future work can explore ways to mitigate unconscious influences and design more secure and fair systems for global, high-stakes assessments.

Acknowledgments

JR Lockwood provided valuable guidance on statistical aspects of the study, and three anonymous reviewers provided feedback that improved the quality of the paper.

References

William Belzak, JR Lockwood, and Yigal Attali. Measuring variability in proctor decision making on high-stakes assessments: Improving test security in the digital age. *Educational Measurement: Issues and Practice*, 43(1):52–65, 2024.

Marilynn B Brewer. The psychology of prejudice: Ingroup love and outgroup hate? *Journal of social issues*, 55(3):429–444, 1999.

Elizabeth Bruch and Fred Feinberg. Decision-making processes in social contexts. *Annual review of sociology*, 43(1):207–227, 2017.

Jill Burstein. The duolingo english test responsible ai standards., 2023. URL https://go.duolingo.com/ResponsibleAI.

Ramsey Cardwell, Benjaimin Naismith, Geoffrey T LaFlair, and Steven Nydick. Duolingo english test: Technical manual. *Duolingo Research Report*, 2024.

Gregory J Cizek. *Cheating on tests: How to do it, detect it, and prevent it.* Routledge, 1999.

Simon Coghlan, Tim Miller, and Jeannie Paterson. Good proctor or "big brother"? ethics of online exam supervision technologies. *Philosophy & Technology*, 34(4):1581–1606, 2021.

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

Phillip Dawson. Remote proctoring: Understanding the debate. In *Second handbook of academic integrity*, pages 1511–1526. Springer, 2024.

William B Gudykunst, Yuko Matsumoto, Stella Ting-Toomey, Tsukasa Nishida, Kwangsu Kim, and Sam Heyman. The influence of cultural individualism-collectivism, self construals, and individual values on communication styles across cultures. *Human communication research*, 22(4):510–543, 1996.

Miles Hewstone, Mark Rubin, and Hazel Willis. Intergroup bias. *Annual review of psychology*, 53(1):575–604, 2002.

Daniel R Isbell, Benjamin Kremmel, and Jieun Kim. Remote proctoring in language testing: Implications for fairness and justice. *Language Assessment Quarterly*, 20(4-5):469–487, 2023.

Manqian Liao, Sinon Tan, and Baig Basim. Plagiarism detection using human-in-the-loop ai., 2023.

Joke Meeus, Bart Duriez, Norbert Vanbeselaere, and Filip Boen. The role of national identity representation in the relation between in-group identification and out-group derogation: Ethnic versus civic representation. *British journal of social psychology*, 49(2): 305–320, 2010.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54 (6):1–35, 2021.

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

Chenhao Niu, Kevin P. Yancey, Ruidong Liu, Mirza Basim Baig, André Kenji Horie, and James Sharpnack. Detecting LLM-assisted cheating on open-ended writing tasks on language proficiency tests. In Franck Dernoncourt, Daniel Preoţiuc-Pietro, and Anastasia Shimorina, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 940–953, Miami, Florida, US, November 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.emnlp-industry.70.

Vinodkumar Prabhakaran, Rida Qadri, and Ben Hutchinson. Cultural incongruencies in artificial intelligence. *arXiv preprint arXiv:2211.13069*, 2022.

PL Patrick Rau, Ye Li, and Dingjun Li. Effects of communication style and culture on ability to accept recommendations from robots. *Computers in Human Behavior*, 25(2): 587–595, 2009.

Yong-Siang Shih, Zach Zhao, Chenhao Niu, Bruce Iberg, James Sharpnack, and Mirza Basim Baig. Ai-assisted gaze detection for proctoring online exams. *arXiv*, 2024. URL https://arxiv.org/pdf/2409.16923.

Paula Winke, Susan Gass, and Carol Myford. Raters' l2 background as a potential source of bias in rating oral performance. *Language testing*, 30(2):231–252, 2013.