Supporting Tutors in the Gig Economy with Automated Feedback: A Case Study on Ringle

Yeon Su Park[1,†], Sieun Kim[2,†], Keighley Overbay[3], Seoyoung Kim[1], Sewook Wee[4], Daho Jung[1] and Juho Kim[1]

[1]KAIST, Daejeon, Republic of Korea
[2]University of Michigan, Ann Arbor, Michigan, USA
[3]Samsung Research, Seoul, Republic of Korea
[4]Ringle, San Mateo, California, USA

Abstract

The rise of online tutoring platforms in the gig economy has made education more scalable, flexible, and on-demand. These platforms rely on learner evaluations as the primary feedback for tutors and platforms. However, such feedback offers limited guidance for tutors' improvement and makes it difficult to monitor tutor quality at scale. To this end, we explored AI-powered automated feedback and how tutors perceive and respond to it. We deployed a research probe on Ringle, a popular online English tutoring platform, that analyzed tutors' lessons and provided automated feedback. We then surveyed 36 tutors about their experience. Our findings reveal that while tutors perceived automated feedback more negatively than learner feedback, they found it useful for self-monitoring and understanding platform expectations, though discrepancies between them often caused confusion. Based on these insights, we propose design considerations for feedback systems for online educational gig platforms.

Keywords

Gig Economy, Online Tutoring Platform, Tutor Support, Language Learning, Automated Feedback

1. Introduction

The gig economy has expanded opportunities for flexible work across various fields. In education, online language tutoring platforms such as Preply, italki, Cambly, Verbling, and Ringle have become increasingly prevalent. Driven by global demand for English learning, these platforms let individuals tutor using their English proficiency, providing scalable and accessible learning [1].

A central challenge of such platforms is maintaining worker quality at scale [2]. To address this, platforms commonly rely on user feedback such as ratings and reviews to evaluate performance and guide improvement. However, prior work shows that such feedback tends to be overly positive, as dissatisfied users often choose not to leave it [3]. As a result, it can be ambiguous for workers seeking concrete guidance and insufficient for platforms monitoring quality.

These limitations are amplified in educational gig platforms. Learner feedback is often treated as a proxy for instructional quality, yet learner satisfaction does not always align with learning effectiveness. Meaningful learning often involves cognitive challenge, which can temporarily reduce satisfaction despite positive outcomes [4]. Thus, learner feedback alone may fail to give tutors actionable insights while limiting platforms' ability to maintain instructional standards.

To address these challenges, we explore AI-powered automated feedback as a complement. Since automated feedback is interpreted alongside high-stakes learner feedback in practice [2], we investigate how tutors perceive and respond to it under this multisource context. We developed an automated feedback system as a research probe and deployed it on Ringle, a popular online English tutoring platform. The system analyzed tutors' lessons using the platform's internal evaluation standards and provided standardized, numerical feedback.

The 2026 Workshop on Pedagogical Evaluation of Automated Feedback (PEAF2026)
[†]These authors contributed equally.

Our survey of 36 tutors found that, although tutors generally perceived automated feedback more negatively than learner feedback, they still found it valuable for self-monitoring and understanding platform expectations. Yet, conflicts with more lenient and high-stakes learner feedback caused confusion and frustration. Based on these findings, we propose design considerations for automated feedback systems in educational gig platforms at scale.

2. Methods

2.1. Study Context

Ringle[1] is a popular online English tutoring platform offering 1-on-1 lessons with native English tutors. Since effectively leading tutoring sessions takes time to learn, tutors are trained through guidance documents and must pass Ringle's internal evaluation by conducting a mock lesson. After each lesson, Ringle encourages learners to leave feedback in three main types. First, learners can leave a 5-star rating on the lesson. Second, they can rate their willingness to meet the tutor again on a 3-point Likert scale. Lastly, they can leave a free-form review of the overall experience. The platform informs learners that 5-star ratings and free-form reviews are shared with the tutor (as feedback) and other learners (as reviews), while willingness-to-meet-again ratings are kept private and shared only with the platform so learners can express honest opinions.

2.2. Research Probe

We collaborated with three Ringle team members to develop automated feedback based on the platform's evaluation standards, previously used in mock lessons to familiarize tutors with internal expectations. These standards consisted of nine pedagogical categories (e.g., lesson structure, engagement), each assessed on a 5-point scale. To examine how tutors perceive standardized automated judgment, we developed an AI-powered feedback system as a research probe that prioritized transparency over complexity, reporting numerical scores for each of the nine subcategories generated from transcripts and tutors' notes [5]. We used existing models with few-shot prompting [6], and score thresholds were iteratively refined with the Ringle team until reaching satisfactory agreement with senior evaluators' manual ratings. Tutors received three reports during their first ten lessons, after the first, fifth, and tenth.

2.3. Survey

2.3.1. Participants

We launched the research probe to newly onboarded tutors[2] on Ringle to examine perceptions of the automated feedback system without prior influence from learner feedback. After each tutor completed their first ten lessons, we distributed the survey. Participation was voluntary, and we excluded tutors who had not reviewed all ten learner feedback records or three automated feedback reports. A total of 36 tutors (ages 18–32; M = 23.28, SD = 3.41) participated. The survey took about 15 minutes, and each participant received a 10 USD Amazon gift card.

2.3.2. Survey Questions and Analysis

The survey had two sections. The first examined tutors' perceptions of automated feedback, with items adapted from prior work on evaluation systems [7, 8]. The same questions were asked for both automated and learner feedback, enabling direct comparison across seven dimensions: understanding, accuracy, fairness, favorableness, evaluator qualification, feedback uptake, and impact. Responses used a 7-point Likert scale (1=strongly disagree, 7=strongly agree). The second

[1] https://www.ringleplus.com/

[2] Tutors who joined Ringle within two weeks of the launch of the probe

consisted of open-ended questions on tutors' challenges at Ringle, perceptions of subcategories, and suggestions for improvement. The full survey is available in the OSF Appendix[3].

To compare tutors' perceptions between the two feedback sources, we conducted Wilcoxon signed-rank tests on the Likert responses. For open-ended responses, we conducted an inductive thematic analysis following the method of Braun and Clarke [9].

2.4. Feedback Data

We gathered automated feedback from the first ten lessons of all 36 participants, including 327 of 360 lessons (33 were discarded due to recording issues). We also collected learner feedback from 10,000 randomly sampled lessons by 6,256 learners over a one-month period within the study duration. We then compared score distributions across the two sources to identify differences in how tutors were evaluated by the AI-powered automation system versus learners.

3. Result

3.1. Perceptions of Automated Feedback

As shown in Figure 1, tutors perceived both learner and automated feedback positively overall, with average ratings above 4 out of 7 across most dimensions, except for evaluator qualification for automated feedback. However, tutors rated automated feedback significantly more negatively than learner feedback across six dimensions: understanding, accuracy, fairness, favorableness, evaluator qualification, and impact. However, feedback uptake showed no significant difference.

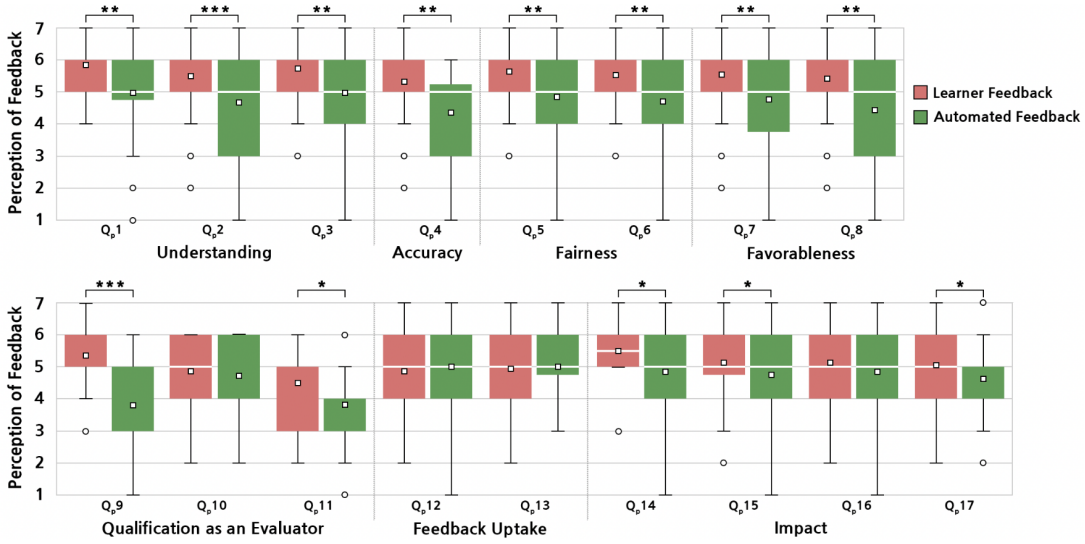

Figure 1: Tutors' perceptions of learner and automated feedback across understanding, accuracy, fairness, favorableness, qualification as an evaluator, feedback uptake, and impact ($^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$)

From open-ended responses, tutors valued the automated feedback for providing clear guidelines on the gig platform's expectations, which helped them better understand performance standards (P3, P20, P19, P24, P28). They perceived this clarity as particularly useful for meeting platform requirements and adapting their teaching to align with structured evaluation standards.

3.2. Differences Between Automated and Learner Feedback

As shown in Figure 2, a substantial portion of lessons received no learner feedback, and when provided, it was highly skewed toward positive ratings (Figure 3, left). In contrast, automated

[3]https://osf.io/zsc3b/overview?view_only=dcf09d1a9fe64765b013df02af6bc382

feedback was consistently available and showed a broader, more fine-grained distribution, capturing greater variation in quality than discrete learner ratings (Figure 3, right).

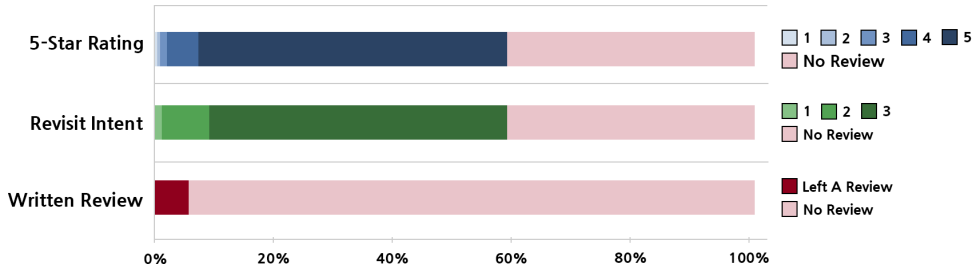

Figure 2: Proportion of learner feedback on 5-star ratings, revisit intent, and written reviews

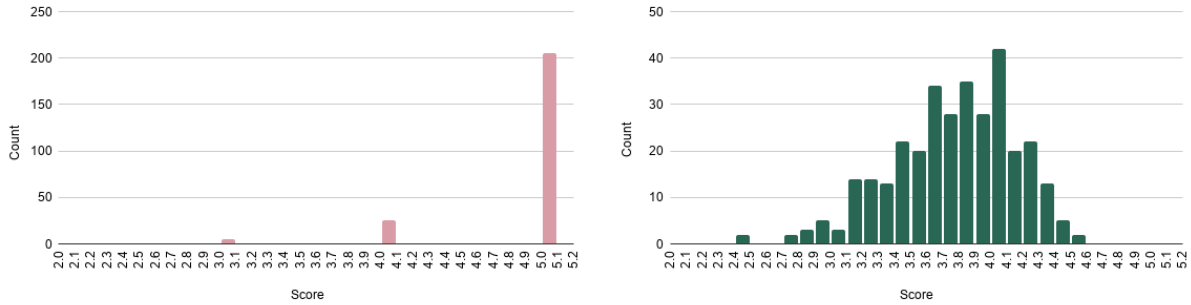

Figure 3: Histogram of 5-level scores from learner (left) and automated feedback (right)

3.3. Responses to Feedback Differences

Confusion Over Score Discrepancies. Many tutors reported confusion over discrepancies between the two feedback sources, particularly because both used comparable 5-point scales (P13, P18, P19, P27, P32, P34). Higher learner ratings (Figure 3) often led tutors to distrust automated feedback, consistent with prior findings on preferences for positive feedback [10].

Contextual Gaps. Many tutors preferred learner feedback because they felt the automated system could not account for varying lesson dynamics (P4, P7, P25, P32, P35). Because the automated feedback lacked contextual awareness—such as learner requests, prior interactions, and time constraints—tutors perceived it as subjective. In contrast, learner feedback was favored as it came directly from the learners and was consequential in a high-stakes evaluative context.

Seeking Rationale. Tutors wanted clearer explanations to justify discrepancies between learner and automated feedback, since automated feedback only presented numerical scores without context (P3, P18, P33). This led them to feel that automated feedback did not sufficiently reflect lesson contexts (P2, P32, P34, P36) and evaluations were inconsistent (P2, P18, P24).

Self-Monitoring. Since learner feedback was highly skewed toward positive ratings, the automated feedback became a valuable tool for self-monitoring and growth. Tutors described using the automated feedback as a self-monitoring tool to identify areas for improvement, track their progress, and gain reassurance about their teaching effectiveness (P3, P35, P36).

4. Design Considerations

We discuss considerations for designing feedback systems in educational gig platforms.

DC1: Distinguish Feedback Types. When multiple feedback sources coexist in settings where learner feedback carries strong evaluative weight [2], it is important to distinguish their purposes and scales. In our study, learner feedback tended to be positively skewed and was often missing altogether (§3.2), limiting its usefulness as an evaluation tool. In contrast, automated feedback provided more consistent and critical assessments that supported tutors' self-monitoring (§3.3). However, presenting feedback sources using similar scoring schemes led to confusion, creating a false impression of equivalent evaluations (§3.3). Explicitly clarifying what each feedback measures and how to interpret the scores can help tutors make sense of multisource feedback.

DC2: Emphasize Complementary Nature. Learner and automated feedback should be framed as complementary rather than conflicting. While tutors often perceived automated feedback negatively—particularly when it failed to capture the lesson context (§3.3)—they found both types of feedback similarly helpful for planning future lessons (§3.1). Discrepancies between feedback sources should therefore be presented as opportunities for reflection rather than signals to prioritize one source over another. Communicating which aspects are better assessed by learners (e.g., engagement) and which are more reliably captured by automated systems (e.g., lesson structure) can support tutors' interpretation and use of feedback for improvement.

DC3: Make Quality Standards Interpretable. Automated feedback can support gig platforms' quality control by organizing evaluations around structured criteria, rather than relying solely on learners' individual impressions. These criteria can make instructional standards more explicit and help tutors understand what the platform expects them to meet (§3.1). To make these standards actionable, systems should pair scores with brief rationales, evidence, or contextual cues that explain how tutors can improve within platform expectations (§3.3).

5. Future Work

Future work should examine how tutors' interpretations of automated feedback evolve over time, particularly as they gain experience navigating multiple feedback sources with unequal stakes. As automated feedback may function as a self-improvement tool rather than merely an evaluative signal, future studies could investigate how tutors adjust their teaching practices in response to such feedback and whether these adjustments lead to measurable improvements in lesson quality. It should also explore whether additional contextual or transparency cues meaningfully change how tutors weigh automated feedback relative to learner feedback in their teaching practices.

Beyond tutors' perceptions, future work should examine the relationship between learner and automated feedback. This could include analyzing whether the two sources are directionally aligned when applied to the same lessons and how strongly they converge or diverge across feedback dimensions. Building on this analysis, future work could explore whether automated feedback can help calibrate positively skewed learner feedback. Understanding these relationships could inform the design of more interpretable feedback systems that help tutors make sense of multiple feedback sources while clarifying what each source is intended to capture.

Declaration on Generative AI

In the preparation of this work, the authors used ChatGPT and Claude for grammar checking and improving the clarity of author-written text. The authors reviewed and revised all content and take full responsibility for the publication's content.

References

[1] M. Xia, Y. Zhao, M. H. Erol, J. Hong, J. Kim, Understanding distributed tutorship in online language tutoring, in: LAK22: 12th International Learning Analytics and Knowledge Conference, LAK22, Association for Computing Machinery, New York, NY, USA, 2022, p. 164174. doi:10.1145/3506860.3506883.

[2] A. J. Wood, M. Graham, V. Lehdonvirta, I. Hjorth, Good gig, bad gig: Autonomy and algorithmic control in the global gig economy, Work, Employment and Society 33 (2019) 56–75. doi:10.1177/0950017018785616, pMID: 30886460.

[3] K. Park, M. Cha, E. Rhim, Positivity bias in customer satisfaction ratings, in: Companion Proceedings of the The Web Conference 2018, WWW '18, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 2018, p. 631638. doi:10.1145/3184558.3186579.

[4] S. K. Carpenter, A. E. Witherby, S. K. Tauber, On students' (mis)judgments of learning and teaching effectiveness, Journal of Applied Research in Memory and Cognition 9 (2020) 137–151. doi:https://doi.org/10.1016/j.jarmac.2019.12.009.

[5] J. Vitale, K. B. Kocabagli, S. Sanghi, A. Van Camp, S. Miller, B. Coker, From noisy classroom transcripts to actionable feedback: Fine-tuning gpt-4o to detect teachers opportunities to respond, in: International Conference on Artificial Intelligence in Education, Springer, 2025, pp. 149–162. doi:https://doi.org/10.1007/978-3-031-98417-4_11.

[6] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners 33 (2020) 1877–1901.

[7] A. Milanowski, H. Heneman, Assessment of teacher reactions to a standards-based teacher evaluation system: A pilot study*, Journal of Personnel Evaluation in Education 15 (2001) 193–212. doi:10.1023/A:1012752725765.

[8] S. Kelly, G. Guner, N. Hunkins, S. K. DMello, High school english teachers reflect on their talk: A study of response to automated feedback with the teacher talk tool, International Journal of Artificial Intelligence in Education 35 (2025) 879–913. doi:https://doi.org/10.1007/s40593-024-00417-x.

[9] V. Braun, V. Clarke, Using thematic analysis in psychology, Qualitative research in psychology 3 (2006) 77–101. doi:https://doi.org/10.1191/1478088706qp063oa.

[10] J. Brett, L. Atwater, 360° feedback: Accuracy, reactions, and perceptions of usefulness, Journal of Applied Psychology 86 (2001) 930–942. doi:10.1037/0021-9010.86.5.930.