Contemporary Educational Technology, 2026, **18**(3), ep674
ISSN: 1309-517X (Online)

OPEN ACCESS

Research Article

AI-augmented judgement in teacher performance assessment: Evidence from a human-AI moderation workflow

Zara Ersozlu [1*]
0000-0002-9120-2921

Susan Ledger [1]
0000-0001-7050-1001

Mark Babic [1]
0009-0002-4635-4405

Robert Parkes [1]
0000-0002-3349-7805

[1] School of Education, University of Newcastle, Callaghan, NSW, AUSTRALIA
[*] Corresponding author: zara.ersozlu@newcastle.edu.au

Citation: Ersozlu, Z., Ledger, S., Babic, M., & Parkes, R. (2026). AI-augmented judgement in teacher performance assessment: Evidence from a human-AI moderation workflow. *Contemporary Educational Technology, 18*(3), Article ep674. https://doi.org/10.30935/cedtech/18970

ARTICLE INFO

Received: 25 Apr 2026

Accepted: 7 Jul 2026

ABSTRACT

Artificial intelligence (AI) is increasingly being explored in educational assessment, but its use in high-stakes performance contexts requires careful design to support quality, consistency, and human accountability. This study designed and evaluated an AI-supported workflow to assist assessors in marking teaching performance assessment (TPA) portfolios. TPA assessors first described how they usually evaluate portfolios, and this process was used to develop a meta-level rubric logic, prompts, and workflow for the AI agent. The workflow was tested by comparing human-only marking with AI supported marking in relation to time, accuracy, cognitive load, feedback quality, usability, bias, and fairness. The findings showed that the AI-supported workflow reduced marking time and perceived workload while supporting more structured, evidence-based feedback. It also contributed to greater consistency and fairness in the moderation process. The AI agent is positioned as a "third eye" that supports human judgement. Human assessors remain responsible for making final decisions and may accept, adapt or override AI-generated suggestions.

Keywords: GenAI, human-in-the-loop, cognitive load, time efficiency, judgement support, moderation, teaching performance assessment

INTRODUCTION

High-stakes professional assessments depend on human judgement where assessors interpret complex evidence, apply standards consistently, and communicate decisions through defensible, criterion-referenced feedback. As a high-stakes and capstone assessment task, teaching performance assessments (TPAs) have different versions and models across different countries and systems. Commonly, teacher candidates are required to provide real-world evidence, such as lesson plans, videos, student work, to demonstrate professional competence. The United States has a highly decentralized ecosystem of TPAs used to evaluate pre-service teachers' readiness for licensure. Their form, governance and uptake vary by state and program, including the national commercial model edTPA, state-based alternatives such as California TPA, the praxis performance assessment for teachers, and local TPAs. In New Zealand TPA's are informed by Australian and US models and typically align with the Teaching Council of New Zealand requirements. Australia has adopted

a nationally mandated TPA assessment, which requires all teacher candidates to complete a TPA in the final year of their preparation program, with evidence annotated against the Australian professional standards for teachers at graduate level (Australian Institute for Teaching and School Leadership, 2011). Despite these system-level and structural differences, all TPAs rely heavily on human judgement and moderation practices to ensure defensible decisions, making them similarly vulnerable to assessor workload, cognitive demand, and variability in judgement. These shared challenges position TPAs across contexts as a compelling site for investigating artificial intelligence (AI)-augmented judgement support that enhances consistency and feedback quality while preserving human accountability.

A recent systematic review of TPAs highlights the difficulties in maintaining consistent marking expectations in high-stakes teacher evaluations given differences in professional standards assessed, rubrics, school contexts in the moderation process (Babic et al., 2026). Assessors' marking and feedback process become cognitively demanding as they work with large and artifact-dense portfolios when evaluating TPAs. Assessors need to be able to coordinate multiple criteria in working memory, search for evidence, make distinctions between performance levels, score fairly and craft feedback that is criterion-referenced, evidence-based and timely. Assessing such large portfolios can sometimes lead to unfair, inconsistent judgement and feedback, limited feedback specificity and reduced transparency, risks well documented in performance assessment research (Bloxham & Boyd, 2012; Mascadri et al., 2023; Stacey et al., 2020).

However, while existing research identifies these challenges in TPA assessment and moderation, less is known about how AI-supported workflows might assist assessors without replacing their professional judgement. This represents an important gap for teacher education, where assessment decisions must be both efficient and educationally defensible. Human moderation of TPAs could benefit from the affordances of generative AI (GenAI). It is commonly used for assessment automation purposes, but this use is often poorly aligned with professional performance assessment, because judgement must remain accountable and contextual (Fadel et al., 2019; Lucking et al., 2016; Williamson & Eynon, 2020). However, considerable risk centers around using only GenAI or automated systems as the sole basis for assessment decisions. The recent Robodebt case in Australia, although not a GenAI case, illustrates the broader risk of relying on automated decision-making systems without sufficient human oversight, transparency, and accountability. A more plausible direction is augmentation, in which AI is designed to support assessors by reducing workload while preserving human authority over final judgements (Corbin et al., 2025; Holmes et al., 2022a; Swiecki et al., 2022). Augmentation can also support the moderation process by providing a "third eye" that makes disagreements and agreements visible and supports assessors to articulate evidence against to professional standards.

The central challenge is not whether AI can generate plausible text, but whether AI-supported assessment workflows can improve judgement quality while maintaining validity, fairness and accountability. In high-stakes assessment, validation depends on standards-based, evidence-based and accountable processes. This shifts assessment priorities from just accuracy alone to socio-technical questions such as what parts of assessor's work can be responsibly supported, what new risks or challenges emerge and what governance practices help ensure that AI-agent systems strengthen rather than weaken fairness and accountability. This study therefore contributes to the field of education by examining AI not as an automated marker, but as a judgement-support tool within a high-stakes teacher education assessment context.

Performance assessment is also a communicative practice. Feedback is the mechanism through which candidates understand expectations and institutions demonstrate procedural fairness. However, feedback quality is vulnerable to human factors such as under-citing evidence, omitting criteria, or relying on generic phrasing when time is constrained (Boud & Molloy, 2013; Hattie & Timperley, 2007). An AI-agent-based judgement-support system can support these processes by making evidence easier to locate and by prompting more consistent use of rubric criteria. This may improve not only efficiency but also the auditability and actionability of feedback. In this sense, the significance of the study lies in its focus on how AI can support better human assessment practices.

This study investigates what changes when AI is introduced as a tool and assistant to TPA scoring, feedback drafting, evidencing, moderating, judgement calibration. It examines how these changes relate to validity, appropriate reliance, feedback quality, and governance in high-stakes teacher performance assessment. We

developed and tested an evaluator-facing account of how AI reshapes assessor cognition and moderation practice in high-stakes teacher performance assessment. By doing so, the study addresses a gap between general discussions of AI in assessment and the practical realities of assessor workload, evidence interpretation, and moderation in TPAs.

We used a design and evaluation approach to develop and test an AI-augmented TPA workflow through a structured workshop. The AI agent system generated rubric-referenced draft feedback organized by standard/criterion and grounded in cited submission evidence where assessors remained responsible for editing and making final decisions. We particularly investigated four areas

(a) cognitive load demands and time efficiency,

(b) feedback structure and specificity,

(c) assessors' perceptions of usefulness, fairness, and appropriateness, and

(d) AI-human convergence/divergence patterns and their use in moderation dialogue.

To avoid over-interpretation, we note that baseline task A and AI-supported task B conditions used different anonymized but similar submissions. The findings indicate plausible magnitudes and mechanisms rather than causal effects attributable to AI support. More specifically, this study investigated:

1. How does an AI agent system impact assessors' perception of their cognitive load and time efficiency?
2. How does the AI agent system influence the structure, specificity, and consistency of feedback?
3. How do assessors perceive the usefulness, fairness, and appropriateness of AI support within moderation?
4. What patterns emerge between AI-generated and human-generated feedback?

LITERATURE REVIEW AND CONCEPTUAL FRAMING

High-stakes performance assessment research shows that assessment quality depends not only on the rubric, but also how assessors interpret evidence, apply standards, justify decisions, and participate in moderation. In teacher performance assessment, these issues are especially important because assessors evaluate complex portfolios. Previous assessment studies have raised concerns about consistency, assessor workload, feedback specificity, and the difficulty of maintaining shared standards across assessors and contexts (Bloxham & Boyd, 2012; Bloxham et al., 2016; Klenowski & Adie, 2009; Mascadri et al., 2023; Sadler, 2013; Stacey et al., 2020). These challenges suggest that TPA assessment is not simply a scoring task, but a demanding professional judgement process that requires evidence interpretation, moderation, and defensible feedback.

Despite growing interest in GenAI in education, much of the existing work has focused on automated grading, predictive analytics, or general feedback generation. Less is known about how AI can support assessor cognition, evidence use, moderation dialogue, and professional judgement in high-stakes teacher performance assessment without replacing human authority (Holmes et al., 2022b; Zawacki-Richter et al., 2019). This is the specific gap addressed in this study. It examines whether and how an AI-supported workflow can assist assessors with evidence, rubric alignment, feedback drafting, and moderation while preserving human accountability for final judgements.

In response to this gap, we frame the AI system as a human-in-the-loop judgement-support interface embedded in a high-stakes assessment workflow, rather than as an automated marker. Three strands inform this framing:

(1) moderation as judgement practice,

(2) assessor cognition and evidence-claim-warrant work, and

(3) human-AI decision support including trust calibration and appropriate reliance.

A common theme is that high-stakes assessment is a socio-technical system: reliability and fairness are produced not only by rubrics and training, but also by tools, time, organizational routines, and accountability structures.

Recent studies on TPA and AI-supported assessment and feedback highlight both the promise of GenAI for improving feedback structure, consistency, and efficiency, and the continuing risks associated with validity, fairness, bias, and over-reliance on automated outputs (Dai et al., 2024; Flodén, 2025; Henderson et al., 2025; Usher, 2025; Zacharis & Papadakis, 2025). In TPA contexts, recent work also shows that assessor judgement, evidence interpretation, moderation, and consistency remain important challenges (Mascadri et al., 2023; Weston et al., 2026; Wyatt-Smith et al., 2024). Recent reviews and policy-oriented studies also suggest that the rapid growth of AI in higher education has outpaced empirical work on how AI should be governed in assessment contexts, especially when human judgement and accountability remain central (Crompton & Burke, 2023; Moorhouse et al., 2023).

Moderation as Judgement Practice

Moderation is a standardization process in which assessors calibrate their interpretations of standards/criteria, discuss evidence sufficiency and develop a shared account of quality (Klenowski & Adie, 2009; Sadler, 2013). In performance assessments, moderation is not only a reliability mechanism, but also a process for strengthening professional judgement by making tacit criteria more explicit (Bloxham & Boyd, 2012; Bloxham et al., 2016).

In Australian TPA contexts, research documents meaningful variation in how assessors interpret rubrics and apply standards and reach consistency (Mascadri et al., 2023; Stacey et al., 2020). Moderation serves multiple purposes beyond comparability, including professional learning and accountability (Bloxham et al., 2016). Wyatt-Smith et al. (2024) and Weston et al. (2026) highlight the role TPAs play in shaping teacher education policy and the importance it could have on the profession if its potential is fully realized. They also highlight ongoing concerns about moderation, digital infrastructure and benchmarking. These studies show that the problem is not only whether TPAs are valid in principle, but how assessor judgement is supported in practice. In the current study, moderation is the practical setting in which AI support is examined. The AI-supported workflow was designed to make assessor agreement, disagreement, and evidence use more visible, so that moderation could become a more explicit and evidence-informed process rather than an informal comparison of scores.

Assessor Cognition

Performance portfolio evaluation requires a set of cognitive and interpretive activities including maintaining focus, coordinating working memory across several criteria, and iteratively mapping evidence, annotations, claims and justification for why specific evidence meets rubric descriptors (Joughin, 2008; Sadler, 2013). The assessor role is not only one of classification and analysis, but it also involves weighting evidence and justifying decisions, which requires significant amount of cognitive processing

In complex assessment contexts such as a TPA, a substantial proportion of cognitive effort is necessarily devoted to the demands of the task itself, reflecting what is known as an intrinsic load. However, poorly structured assessment processes can generate considerable extraneous load, such as searching for evidence across multiple documents, switching repeatedly between screens, or drafting evaluative comments. Such activities consume cognitive resources without contributing to the quality of judgement (Paas et al., 2003; Sweller, 1988; Sweller, 2011). For high-stakes evaluative work, it is preferable that assessors' cognitive effort is directed toward interpreting evidence and developing robust judgements and feedback for the submission.

An AI-enabled judgement-support tool can function as an additional cognitive workspace that reshapes how assessors engage with the task. Rather than relying on working memory to hold and organize multiple strands of evidence, the tool can assist in organising, analyzing, and classifying information throughout the assessment process (Risko & Gilbert, 2016; Sweller, 2011). By generating preliminary syntheses or interpretive summaries, it enables assessors to allocate more of their cognitive resources to evaluating the quality of performance and making defensible professional judgements. It is a clear rationale that investigating AI support in TPAs, where the assessment task is cognitively demanding and the consequences of judgement are significant. This concept directly informs the study's focus on assessor workload, marking time, and cognitive load. By examining how assessors use AI to locate evidence, organize rubric-aligned feedback, and reduce repetitive searching, the study evaluates whether AI support can redirect cognitive effort toward professional judgement rather than administrative processing.

Human-in-the-Loop AI as Judgement Support

Discussions on using AI in assessment highlight both opportunities such as scalability, structure, consistency and also risks such as opacity, bias, narrowing of quality (Corbin et al., 2025; Swiecki et al., 2022; Williamson & Eynon, 2020). In high-stakes assessment contexts, a human-in-the-loop approach that sees AI as a decision support tool can preserve accountability while improving assessment quality. AI can be also used as a "third eye" in moderation which supports assessor confidence or prompt articulation and evidence re-checking.

Research on human-AI collaboration reveals that transparency, uncertainty signaling and explainability significantly affect trust and reliance behaviors (Bansal et al., 2021; Hancock et al., 2011). For judgement support indented tools, this implies careful design to workflow features that make evidence visible, make uncertainty legible, and encourage verification when outputs are contestable (Lee & See, 2004; Parasuraman et al., 2000). Research on decision support tools suggest that people use those tools appropriately if the system is clear about how it reached its output and if users can check, question and change the result. In assessment settings, this means designing tools that clearly link judgements to evidence, prompt assessors to look for counter evidence and make editing and revising normal practice. Therefore, AI-supported assessment should not be judged only by whether it saves time, but also by whether it makes decisions more transparent, evidence-based, and traceable. This framing underpins the design of the AI agent in the present study. The system was not designed to replace assessors or make final decisions; instead, it generated evidence-linked, rubric-referenced draft feedback that assessors could accept, revise, or reject.

Automation Bias as a Predictable Risk

Automation bias existed even AI is positioned as assistive or decision support system for human judgment meaning it can influence how people think and decide through over reliance even AI suggestion is wrong (commission errors) or fail to notice mistake because they place too much trust in the AI output (omission errors) (Cummings, 2004; Mosier et al., 1998; Parasuraman & Riley, 1997).

Automation bias is particularly risky in high-stakes assessment, because consequences of errors are serious for all stakeholders who take part. It becomes a problem when assessors accept AI-logic without verifying student's actual evidence, or when AI highlights a certain criterion, human assessor can pay less attention to other criteria. These same risks apply to educational applications of AI, especially where accountability for judgement must remain with human professionals (Holmes et al., 2022a; Skitka et al., 1999).

The risk of automation bias should not be a reason to reject using AI decision support all together. The problem here is not AI itself but how it is implemented, understood and structured. AI must be designed and governed in ways that protect and build on human judgement. Systems built to use AI should force or strongly encourage human verification. To address these risks, the AI-supported workflow in this study required evidence citations or exact excerpts from the submission for each feedback claim, encouraged assessors to actively review AI-generated suggestions, and asked assessors to explain their reasoning when they disagreed with AI output. This process supported metacognition and created an audit trail for human override decisions. We also used calibration checks to maintain appropriate reliance, continued human critical engagement to avoid where people may trust AI without checking over time. In this way, the AI-supported assessment system was designed to encourage assessors to verify evidence, justify decisions, and remain accountable for final judgements. In the current study, automation bias is treated as a design and governance issue. For this reason, the workflow required assessors to verify AI-generated feedback against cited evidence, explain disagreement where relevant, and retain responsibility for final judgement.

RESEARCH METHOD AND DESIGN

We conducted a qualitative dominant mixed methods design and evaluation study using a structured workshop format. The researcher-led workshop enabled comparison of assessor work under two conditions: assessment-as-usual without AI and assessment with AI support. We collected time-on-task, perceived cognitive load, NASA-TLX (Hart & Staveland, 1988) system usability scale (SUS) (Brooke, 1995), reflection responses, moderation discussion data, and assessor feedback texts. Given the small sample (n = 6),

quantitative findings are reported descriptively (means and ranges) to indicate magnitude and direction of change rather than to support inferential claims.

Context

The study was situated in an Australian university TPA program in which final-year pre-service teachers submit portfolios demonstrating teaching capability aligned to the Australian professional standards for teachers at graduate level (Australian Institute for Teaching and School Leadership, 2011). Submissions include a context description, lesson sequence plans and materials, student work samples with assessment/feedback, and reflective commentary. Assessors typically spend 3-4 hours per submission and participate in moderation processes including double-marking and calibration meetings.

Participants

Six teacher educators participated, purposively sampled for variation in TPA assessment experience: two with extensive experience (> 50 TPAs assessed), two with moderate experience (20-49 TPAs assessed), and two with limited experience (< 20 TPAs assessed). Mean TPA assessment experience was 7 years (range: 3-10 years). All participants had school teaching experience (mean: 11 years, range: 6-18 years) and held academic appointments at the institution and/or professional roles connected to school-based teacher education and assessment. Participants were recruited via email invitation to current TPA assessors who had assessed at least 5 portfolios in the previous two years. Response rate was 67% (6 of 9 invited assessors). Participants received 4 hours compensation beyond their existing role responsibilities. Purposive sampling was selected because the study required participants with direct experience of TPA assessment, rubric interpretation, evidence-based judgement, and moderation. The aim was not statistical representativeness, but to obtain information-rich accounts from assessors who could meaningfully evaluate the AI-supported workflow. Variation in assessor experience was also intentionally sought so that the study could explore how different levels of expertise shaped responses to AI augmentation. The inclusion criteria were that participants were current or recent TPA assessors, had assessed at least 5 TPA portfolios in the previous two years, were familiar with the relevant TPA rubric and moderation expectations, and were available to participate in the structured workshop. No participants were excluded on the basis of demographic characteristics. Exclusion criteria were limited to individuals who did not have relevant TPA assessment experience, had not assessed the minimum number of portfolios, were unavailable for the workshop, or did not provide informed consent. These criteria were used to ensure that participants could engage meaningfully in the assessment task and provide informed feedback on the AI-supported workflow.

While the sample size limits inferential statistical analysis, it is appropriate for this design and evaluation pilot study focused on generating rich qualitative insights, identifying mechanisms of change, and informing larger-scale evaluation. The purposive sampling strategy prioritizes variation in experience levels to explore how expertise moderates responses to AI augmentation. Ethics approval was obtained prior to data collection, and all participants provided informed consent before taking part in the study.

AI Augmented Judgement Support Tool: System Design and Protocol

The tool was developed as an AI-supported judgement aid, combining a custom-built interface with an agent architecture built on GPT-4 (**Figure 1**). It was designed specifically to support rubric-referenced evaluation through a more structured analytic process than would be possible with a general conversational system.

In practice, the tool served three main purposes. First, it helped align portfolio content with the relevant NTPA rubric criteria and associated graduate teaching standards. Second, it highlighted specific excerpts from the portfolio that could be checked directly by the assessor, making it easier to verify the basis of any suggested interpretation. Third, it generated draft feedback wording and indicative scoring prompts based on the evidence it had identified.

A deliberately cautious approach guided the development of the tool. It was designed to remain closely anchored to the submitted text, avoid making unsupported assumptions, and signal uncertainty when the evidence was unclear or open to interpretation. The emphasis was not on producing polished or expansive responses, but on supporting consistency, transparency, and traceability in the assessment process. The

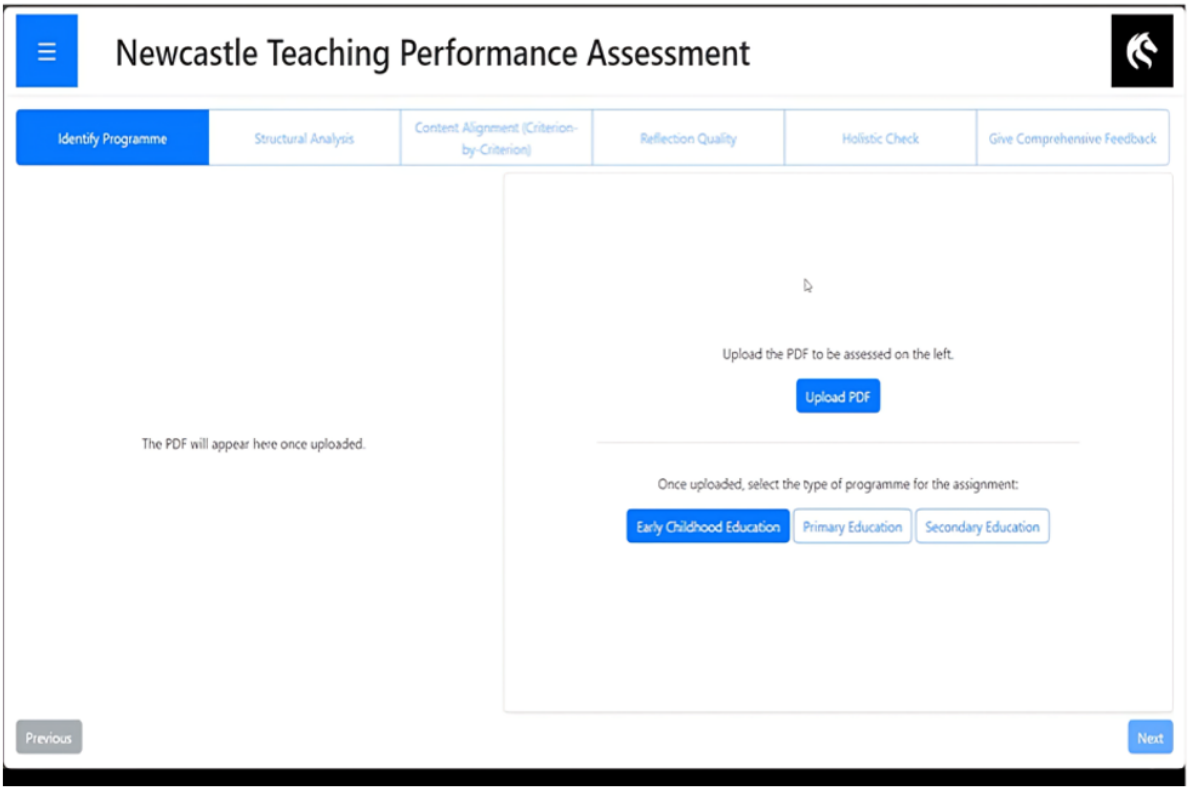

Figure 1. AI augmented judgement support tool (created by the authors)

system was trialed with 12 training portfolios to check that its output remained aligned with established expectations and common assessment practice.

The tool was introduced as a support for professional judgement, not as a substitute for it. Its use was built around a verification-first process intended to keep human decision-making central and to reduce the risk of over-reliance on automated suggestions.

Assessors were asked to form an initial judgement before viewing any system-generated output. This was important in preserving an independent reading of the portfolio and ensuring that the assessor's own professional reasoning came first. Any suggestions produced by the tool were then treated as provisional rather than authoritative. Assessors were expected to check those suggestions against the original portfolio and decide whether they were accurate, useful, or incomplete.

Responsibility for final scores and written feedback remained entirely with the human assessor. In practice, the tool was most useful in drawing attention to possible evidence, supporting consistency in feedback language, and prompting closer review of particular sections of a submission. Where differences emerged between assessor judgement and tool-generated suggestions, these were explored through moderation discussion. In this way, the tool supported professional dialogue and critical reflection, rather than replacing the expertise on which sound judgement depends.

Workshop Procedure

The workshop lasted for 3.5 hours and included pre-work and in-session activities:

- Orientation and calibration (45 minutes): Participants reviewed the study purpose, were reminded of moderation expectations, and completed a familiarization activity with the AI interface and the rubric structure.
- Baseline assessment without AI (90 minutes): Participants assessed one anonymized submission (task A, portfolio focusing on literacy teaching in year 3) using standard procedures, recorded time, and submitted feedback. Participants worked independently without access to the AI system.

- AI-supported assessment (90 minutes): Participants assessed a second anonymized submission (task B, portfolio focusing on mathematics teaching in year 5) using the AI system. They generated AI draft feedback, reviewed it against their reading of the submission, and edited/supplemented as needed while tracking time. A 15-minute break separated task A and task B.
- Reflection and moderation dialogue (45 minutes): Participants completed post-task surveys (NASA-TLX, SUS, open-ended reflection questions) and participated in a facilitated focus group discussion focused on usefulness, fairness, and how AI-human agreement/disagreement could be used within moderation.

To reduce order effects, the baseline task was completed first for all participants, followed by the AI-supported task; this sequencing reflects a pragmatic workshop design rather than an optimized experimental counterbalance. The use of different submissions in each condition was necessary to protect candidate anonymity and avoid ceiling effects from repeated assessment of the same portfolio. While this design prevents direct causal attribution of all differences to AI support alone, the large effect sizes and consistent qualitative patterns suggest AI played a substantive role.

During the moderation dialogue, the facilitator used semi-structured prompts to elicit:

(1) points where AI output aligned with assessors' judgements,

(2) instances of disagreement and how these were resolved,

(3) perceived risks and governance requirements for high stakes use, and

(4) reflections on how the tool might reshape moderation practice.

The discussion was audio-recorded and transcribed.

Data Sources and Analysis

We used both quantitative and qualitative data to respond to research questions. As the quantitative analysis we used:

(1) time-on-task for task A and task B,

(2) NASA-TLX administered after each task,

(3) SUS administered after task B,

(4) open-ended reflection responses,

(5) focus group transcript, and

(6) assessor feedback texts from both conditions.

NASA-TLX was calculated following the raw score method, with scores on six subscales (mental demand, physical demand, temporal demand, performance, effort, and frustration) ranging from 0-100. Overall cognitive load was calculated as the mean across subscales (Hart & Staveland, 1988). SUS scores were calculated using the standard formula, with scores above 68 indicating above-average usability (Brooke, 1995). Time-on-task was recorded via participant self-report using a standardized timer. Given the sample size, all quantitative data are reported descriptively (means, standard deviations, and ranges) without inferential statistics.

As the qualitative analysis we used reflection responses and focus group data which were analyzed using reflexive thematic analysis (Braun & Clarke, 2006). Two researchers independently coded 40% of the data to establish reliability (Cohen's kappa = 0.82), then divided remaining data for coding. Codes were iteratively refined through discussion and organized into themes related to perceived cognitive support, feedback practices, fairness concerns, trust calibration, and moderation value.

Assessor-edited feedback from each condition was examined for the criterion coverage (proportion of rubric standards addressed), presence of sub-headings by standard/criterion, proportion of evidence-linked statements and the frequency of developmental suggestions.

Triangulation was supported by comparing survey responses, time records, and focus group discussion against edited feedback texts to check whether reported experiences were reflected in observable feedback features.

Table 1. NASA-TLX cognitive load scores before and after using the AI system

Dimension	Without AI (M)	With AI (M)	Change	Percentage reduction
Mental demand	90	45	-45	50%
Physical demand	75	35	-40	53%
Temporal demand (time pressure)	70	20	-50	71%
Performance (success)	50	80	+30	+60%
Effort required	80	35	-45	56%
Frustration	50	15	-35	70%

Note. Scores range from 0 (low) to 100 (high) & for performance, higher scores indicate greater perceived success

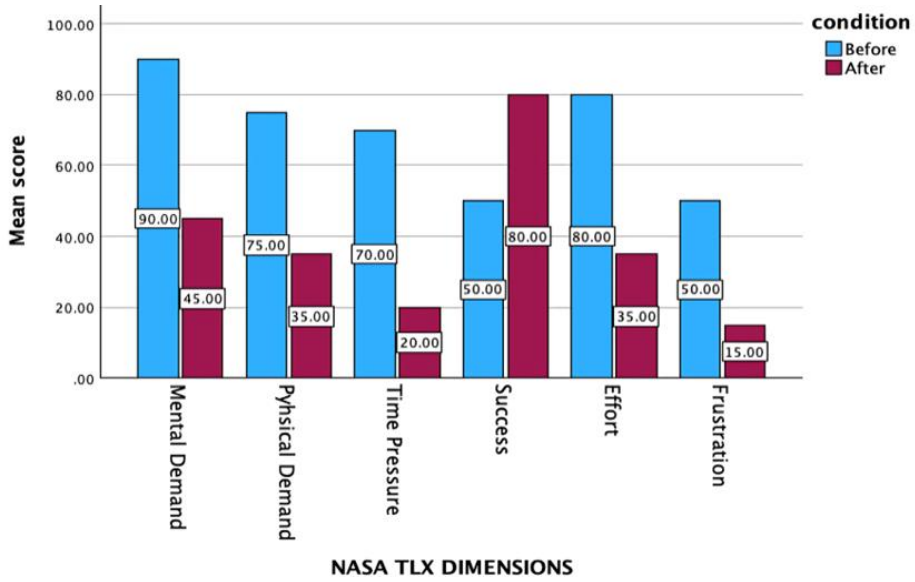

Figure 2. NASA TLX cognitive load before and after using AI tool (the authors' own elaboration)

Table 2. Marking time per assessor (minutes)

Assessor	Task A (without AI)	Task B (with AI)	Time saved	Percentage reduction
Assessor 1	53	~10	43	81%
Assessor 2	90	~15	75	83%
Assessor 3	32	~8	24	75%
Assessor 4	66	~12	54	82%
Assessor 5	57	~10	47	82%
Assessor 6	175	~25	150	86%
Mean	78.8	~13	65.5	82%

Note: Task B times are approximate because participants completed at varying paces & the mean is based on facilitator observations

FINDINGS

Cognitive Load and Time-on-Task

Across participants, perceived cognitive load decreased markedly on NASA-TLX dimensions, with the largest reductions in temporal demand (time pressure), effort, and frustration. Participants described the workflow shift from generating feedback "from scratch" to evaluating and improving a draft, which they perceived as a better use of attention. Time-on-task also decreased substantially in the AI-supported condition; participants attributed this to reduced writing overhead and more systematic criterion coverage rather than to reduced scrutiny of evidence.

Table 1 shows the NASA-TLX cognitive load scores before and after using the AI system. **Figure 2** depicts the NASA TLX cognitive load before and after using AI tool. **Table 2** shows the marking time per assessor (minutes). **Figure 3** shows the marking time per assessor before and after AI-supported workflow use.

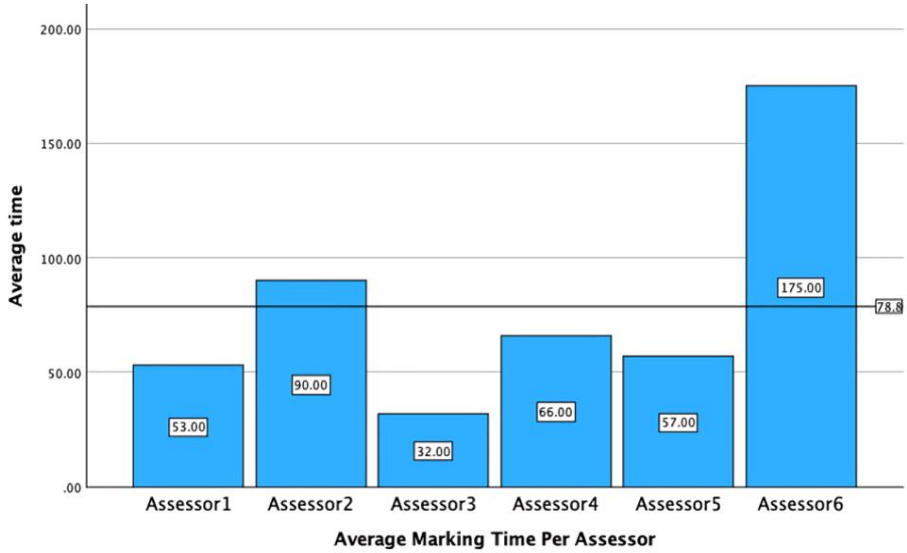

Figure 3. Marking time per assessor before and after AI-supported workflow use (the authors' own elaboration)

Table 3. Feedback structure comparison

Feature	Without AI	With AI
Standards addressed (mean)	5.3 / 7	6.8 / 7
Average words per standard	47	89
Use of sub-headings	2 / 6 assessors	6 / 6 assessors
Evidence citations	33% of feedback	78% of feedback
Developmental suggestions	42% of feedback	91% of feedback

Feedback Structure, Specificity, and Consistency

Feedback produced with AI support was more consistently organized by standard/criterion and more likely to include evidence-referenced statements (**Table 3**). Differences were largest for less experienced assessors, who reported that the AI outputs provided a useful template and "quality checklist." Experienced assessors reported smaller changes but still valued systematic coverage and the ability to reallocate effort toward judgement checking and contextualization.

Participants' annotations and revisions suggested that the AI drafts often improved the surface qualities of feedback (clear headings, consistent structure, and a balance of strengths and next steps), but assessors frequently adjusted the substance of claims to better match context. Common edit patterns included: softening overly definitive language where evidence was thin, adding contextual qualifiers (e.g., references to classroom constraints described in the submission), and tailoring developmental suggestions to be more actionable for the candidate. In several cases, assessors reported that the AI drafts helped them notice a missing element (e.g., a standard not yet addressed) but that the most valuable work remained deciding what counts as sufficient evidence for a particular descriptor.

The consistency gains were also described as a moderation benefit. Several assessors noted that when feedback is structured in a comparable way across submissions, moderation conversations become more efficient because assessors can rapidly locate evidence claims, compare warrant strength, and identify where two assessors are applying different thresholds. This suggests that feedback structure is not merely a communication outcome for candidates but also an infrastructural support for moderation work. **Table 3** compares key feedback features across conditions.

Perceived Usefulness, Fairness, and Appropriateness

Participants completed post-task usefulness items and the SUS. The system achieved a mean SUS score of 87, indicating excellent usability. Post-workshop ratings also indicated high perceived ease/helpfulness (ease of marking with AI: 100%; helpfulness overall: 92%) alongside calibrated confidence in AI accuracy (84%).

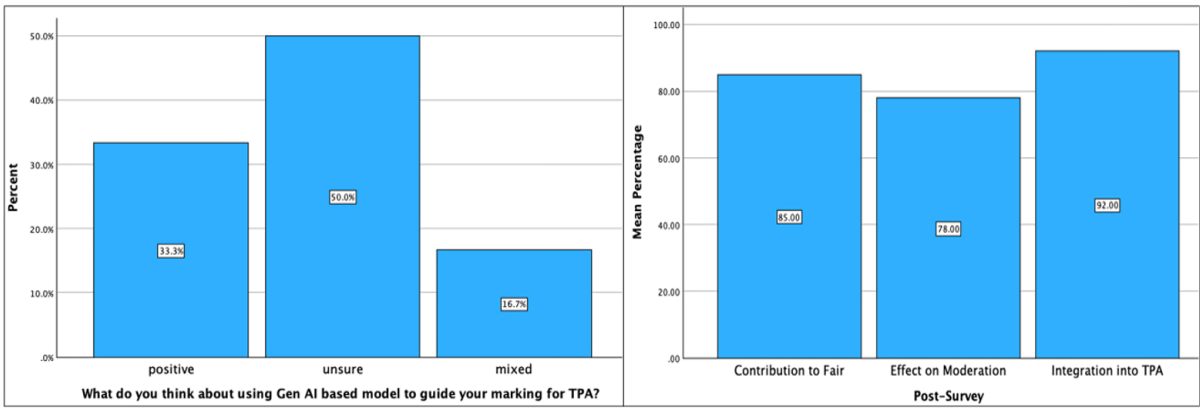

Figure 4. Perceptions on using AI-supported tool (pre-post) (the authors' own elaboration)

Qualitative reflections help explain why usefulness was high despite acknowledged accuracy caveats. Participants described the tool as an "accelerator" for organising comments across multiple criteria, helping them start feedback sooner and reducing time pressure, which supported a more reflective tone. At the same time, several assessors noted the need to "slow down" at key decision points (e.g., borderline levels) because fluent drafts could mask weak warranting; in these cases, usefulness depended on having evidence excerpts immediately available for verification.

Fairness was discussed primarily as consistency of process (criterion coverage and reduced fatigue effects) rather than uniformity of outcomes, and participants explicitly noted that consistency is not equivalent to fairness. They emphasized governance conditions for appropriate use, AI positioned as assistive, explicit expectations for evidence checking, and transparency about known limitations. Participants also identified potential fairness risks if the system under-weights forms of evidence that are difficult to represent textually or over-generalizes from stereotyped exemplars, reinforcing that perceived fairness depends on both representational choices (what the model "sees") and organizational safeguards (what must be checked and what is audited). **Figure 4** shows the perceptions on using AI-supported tool (pre-post).

Convergence/Divergence and Moderation Value

We compared AI-generated and human-generated feedback to identify patterns of agreement and difference. Across coded judgements, strong convergence (similar strengths/weaknesses) accounted for 52%, partial convergence (similar overall judgement with different evidence emphasis) for 26%, divergence (different conclusions) for 15%, and instances where AI surfaced evidence not mentioned by humans for 7%.

Convergence was most common on explicit, documentable criteria, whereas divergence occurred more often when judgements relied on contextual or implicit evidence (e.g., interpreting artifacts beyond transcripts, or weighing thin evidence). Participants treated disagreement as moderation material: divergences prompted assessors to articulate rationales, revisit evidence, and surface tacit calibration rules, with AI described as a "third eye" that made both over-severity and over-generosity more visible.

Three recurring divergence patterns were observed in terms of overconfident generalizations from a small number of excerpts (leading assessors to seek corroboration or temper claims), under-specified interpretations when evidence was distributed across artefacts and required synthesis and threshold differences about what counts as "sufficient" evidence for a descriptor. The 7% "AI-surfaced evidence" instances functioned as coverage checks, either revealing overlooked excerpts or prompting discussion about whether a superficially relevant passage was actually decisive.

DISCUSSION

Ongoing concern about the sustainability of the labor-intensive TPA moderating and assessment process by ITE providers resourcing TPA in their current forms, has resulted in a need to explore alternative options. The findings from this pilot are promising and position AI augmentation less as "speeding up marking" and

more as a redesign of judgement work. Areas of affordances relate to reducing cognitive load of assessors, improving the reliability of feedback, systematically addresses disagreements, and overall improved fairness.

The largest changes occurred in dimensions that plausibly reflect reduced extraneous burden: temporal demand (**Table 1**) and effort decreased substantially (**Figure 2**) and time-on-task dropped markedly (**Table 2** and **Figure 3**). Participants' accounts help interpret these changes: the AI shifted the workflow from blank-page drafting to evaluating and improving a structured draft. From a cognitive load perspective, this is consistent with reducing extraneous load (drafting and searching) to free capacity for the higher-value work of evidence checking and warranting judgements (Risko & Gilbert, 2016; Sweller, 1998, 2011).

AI support was associated with feedback that was more systematically organized and more frequently evidence-referenced (**Table 3** and **Figure 4**). These shifts are consistent with a judgement support interpretation: the system appears to scaffold criterion coverage and evidence citation, thereby strengthening the defensibility of judgements and reducing the likelihood of omission under time pressure within the TPA marking process. The larger changes observed for less experienced assessors suggest that judgement support may reduce variance attributable to assessor expertise differences, at least for feedback structure and coverage.

The convergence/divergence analysis suggests an additional contribution beyond efficiency and feedback structure: disagreement can be a resource for moderation. While reliability framings often treat disagreement as error to be minimized, moderation scholarship treats productive disagreement as central to judgement practice (Klenowski & Adie 2009; Sadler, 2013). Our findings show that, AI-human divergences prompted assessors to critically check, revisit evidence, and surface implicit assumptions. This supports a design implication for human-in-the-loop judgement support rather than using AI to enforce agreement, systems can be designed to surface and structure disagreement. Such design preserves human authority while strengthening the defensibility of decisions.

Finally, fairness and governance remain central (**Figure 4**). Participants linked perceived fairness to consistency and reduced fatigue effects, which is plausible in high-volume assessment contexts. However, consistency is not equivalent to fairness, AI outputs may over-infer from thin evidence, under-handle implicit evidence or reproduce historical patterns embedded in exemplars and rubrics (Green & Chen, 2019; Raji et al., 2020). These risks point to bounded trust and governance requirements: explicit positioning of AI as assistive; mandatory evidence checking; transparency about limitations; and ongoing monitoring for bias and meaning (Corbin, 2005; Swiecki, 2022). Without such guardrails, AI may quietly displace professional deliberation by making judgements feel "settled" before they have been properly tested through evidence and moderation (Skitka et al., 1999).

The magnitude of change observed in this pilot (notably time-on-task and perceived temporal demand) is striking Given the workshop design and the use of different submissions across conditions, these findings should not be read as stable effect sizes. Instead, they provide evidence about plausible mechanisms:

(1) drafting support shifts effort from generation to evaluation (Kellogg, 2008),

(2) systematic criterion coverage reduces omission (Boud & Molly, 2013); and

(3) evidence surfacing supports rapid checking consistent with research on cognitive offloading and external representations (Risko & Gilbert, 2016).

In other words, the pilot offers a credible account of how a judgement-support interface can reallocate assessor cognition in ways that are aligned with defensibility goals.

Two additional mechanisms warrant attention. First, the AI outputs appear to function as a coverage prompt by presenting a full set of rubric-aligned headings, the system reduces the chance that assessors will unintentionally under-attend to a criterion when time is limited (Sweller, 2011). Second, AI drafts can act as a language stabilizer by providing consistent phrasing for rubric constructs reducing variation in how assessors translate standards into narrative feedback (Sadler, 2010). This makes moderation conversations more focused on evidence sufficiency than on wording differences.

These mechanisms are promising, but they also imply that future evaluations should measure not only speed and perceived load, but also whether coverage gains come at the expense of deeper engagement with ambiguous evidence. Methodologically, this pilot points to design requirements for stronger and larger

studies. Subsequent studies should counterbalance submissions across conditions (or use within-submission cross-over designs), include independent ratings of submission difficulty, and pre-register analysis plans for key outcomes (time, cognitive load, feedback quality, and convergence/divergence), consistent with best practices in experimental and educational research design (Ioannidis, 2005). These steps would allow researchers to distinguish workflow gains from artefact difficulty effects and to estimate the conditions under which gains are most reliable (e.g., novice vs. expert assessors; text-heavy vs. screenshots/ pictures heavy portfolios).

A "Third-Eye" Model of Moderation

From an educational technology standpoint, the value of AI may be highest not where it agrees with humans, but where it creates structured opportunities to test warrants. In moderation meetings, disagreement can function like a stress test: it forces assessors to specify what evidence they treated as decisive, what alternative interpretations were rejected, and which rubric descriptors were privileged. This "third-eye" model (AI-human augmentation) reframes AI as a device for increasing the auditability of judgement, not simply for increasing agreement. The practical implication is that interfaces should deliberately support disagreement workflows (flagging, prompts, and documentation) rather than hiding them.

To realize this model in practice, systems should represent divergence at the level of warrants, not only at the level of scores. For example, when the AI suggests a different performance level, the interface can present:

(1) the minimal evidence excerpt(s) that motivated the suggestion,

(2) the rubric descriptor language it is mapping to, and

(3) an explicit "counter-evidence" slot where assessors can paste or note the artefact segment that led them to a different conclusion.

This structure keeps disagreement evidence-centered and reduces the risk that moderation becomes a negotiation of opinions (Sadler, 2010). It also reflects principles from explainable AI, where systems are expected to make the basis of their outputs transparent and open to scrutiny (Miller, 2017).

The "third eye" can also support moderation equity by standardizing which questions are asked of each submission. If every assessor is prompted to consider both confirmatory and dis-confirmatory evidence for each criterion (e.g., "what would move this up a level?", "what is missing for the next level?"), moderation discussion becomes less dependent on individual assessor habits and more anchored to comparable warrant structures. This is consistent with recent research on structured professional judgement and bias reduction, which shows that standardized procedures and explicit consideration of alternative hypotheses can improve consistency and reduce bias in expert decision-making (Oberlader, 2025; Sinha et al., 2025). Importantly, this does not require AI to be correct; it requires the AI to be interrogable and to make warrant structure visible.

Appropriate Reliance in High-Stakes Judgement Support

The usability and confidence findings of AI-Human augmentation in TPA assessing suggest participants experienced the system as easy to use while still maintaining caution about accuracy. This is the desired direction for appropriate reliance: high usability paired with verification norms (Lee & See, 2004). For high-stakes use, appropriate reliance is not a stable trait of the user or a static property of the system; it is produced by design choices (evidence linking, uncertainty cues, and friction for high-confidence claims on thin evidence) and by governance (training, policy, and audit trails). Future evaluation should therefore treat appropriate reliance as a measurable outcome of the socio-technical system, not merely an attitude.

Designing for appropriate reliance also means anticipating where "helpful" drafts can subtly shift the locus of judgement. Participants described valuing structured text because it reduced the blank-page burden; however, the same structure can create inertia (accepting a coherent narrative even when warrants are weak) which aligns with research on automation bias and over-reliance, where users tend to accept system outputs even when evidence is insufficient (Ferrara, 2024; Skitka et al., 1999). Interfaces can counteract this by

(1) visually separating claims from evidence (so that verification is a distinct step),

(2) marking which claims are grounded in direct quotes vs. inferred, and

(3) supporting rapid toggling back to the source artefact at the point of each claim.

These design principles are consistent with work in explainable and human-centered AI, which emphasizes transparency, inspectability, and support for verification over passive acceptance (Miller, 2019; Raji et al., 2020). In moderation settings, brief warrant prompts (one sentence plus an evidence pointer) can provide enough friction to keep the final judgement owned by the assessor while still benefiting from drafting support.

Appropriate reliance should be treated as a collective property of moderation practice, not just an individual behavior. In high-stakes programs, such as TPAs reliance norms are shaped by shared templates, workload expectations, and what is audited. Establishing "check-first" routines (e.g., verifying at least one cited excerpt per criterion; documenting reasons for accepting AI-suggested level shifts) can make reliance visible and coachable, and can help programs detect drift early (e.g., rising rates of verbatim acceptance or decreasing diversity of evidence cited). This reflects broader findings that human-AI reliance is socially and organizationally situated rather than purely individual (Liao & Varshney, 2021).

Our findings show that TPA assessors are less likely to over-trust the AI when the tool is easy to use but they remain cautious about its accuracy. Li et al. (2024) and Ilieva et al. (2025) similarly reported on accuracy concerns which require human oversight when using AI for consistent and accurate marking. Ease of use does not fully erase caution by assessors, and we agree with previous literature that structured moderation practices may help to counteract automation bias.

The adoption of AI for assessments without appropriate pedagogical framing or regulation creates risks to quality and should be adopted with clear structures that echo our findings that efficiency can create marking bias (Li et al., 2024). Bias can occur because the AI reduces marking time and produces convincing drafts, which may encourage "good-enough" acceptance without careful cross-checking. This risk is well documented in studies of automation bias and over-reliance, where users tend to accept system outputs even when they are incorrect or insufficiently supported (Parasuraman & Riley, 1997; Skitka et al., 1999) The divergence discussions are important in this respect, when disagreement with the AI is treated as a normal part of TPA moderation, assessors are required to explain why they accept or reject AI suggestions, therefore reducing passive acceptance (Miller, 2019). In future deployments, it would be useful to monitor simple indicators of verification behavior such as how often assessors edit AI-generated text, how much is accepted unchanged, and whether referenced evidence is actively cross-checked. This aligns with emerging research on human-AI collaboration and auditing, which emphasizes the importance of tracking reliance patterns and verification practices to detect over-trust and performance drift (Green & Chen, 2019; Lee, 2024; Raji et al., 2020). Structured processes, transparency and verification methods are needed by assessors to prevent passive acceptance of AI suggestions.

Based on our pilot, five design implications emerge for AI-supported TPA judgement systems. First, interfaces should foreground evidence visibility by requiring cited excerpts for each feedback claim, enabling rapid warrant verification. Second, divergence between AI and human judgement should be treated as a productive signal rather than an error state, with workflows prompting articulation of reasons for acceptance or rejection. Third, trust calibration must be supported through explicit limitation statements and contestable outputs. Fourth, workflow integration should shift effort from blank-page drafting to evaluative refinement. Finally, governance structures must preserve human accountability through policy clarity, auditability, and monitoring for bias or drift.

Limitations, Implications, and Future Research

This evaluation was conducted as a single workshop with a small sample from one institution and should be interpreted as a pilot study designed to generate design insights and hypotheses for later field deployment. The workshop format prioritized comparison and moderated discussion over ecological validity; longitudinal in-situ studies are required. In addition, the assessment-as-usual and AI-supported conditions used different anonymized submissions, so observed differences should be interpreted as plausible mechanisms and indicative magnitudes rather than causal effects attributable solely to AI support.

The AI system also introduces boundary conditions. Outputs are sensitive to prompt design, rubric representation, and the evidentiary surface available to the model. These constraints shape what the system can cite and increase the risk of over-inference when evidence is thin or implicit. Finally, this pilot focused on

assessor experience and workflow metrics; it did not directly assess downstream consequences for candidates (e.g., uptake of feedback) or institutional effects over time (e.g., moderation culture and drift).

Despite these limitations, the pilot points clear design and governance priorities: evidence-linked feedback to support warrant checking; support for productive moderation by treating AI-human divergence as a calibration input; and interface cues that promote appropriate reliance, including rapid cross-checking and easy revision or rejection of AI text. Future research should test these principles in longitudinal deployments using controlled designs that separate submission difficulty from AI effects, behavioral indicators of verification (e.g., edit patterns and verbatim acceptance), candidate-centered outcomes (usefulness and equity of feedback), and fairness/robustness audits across candidate groups, assessors, and portfolio formats.

CONCLUSION

The findings from this pilot study provide a cautiously hopeful picture of the potential benefits of AI augmentation to complex, high-stakes human judgement required work.

The findings indicated that AI's contribution lies in reorganizing the cognitive architecture of how experts make and convey judgements. AI support seems to free up human capacity for evidence checking and justification testing for defensible judgement. Framing this support as from automation to augmentation has meaningful implications for how AI integration should be conceptualised, designed and evaluated across professional judgement required contexts.

An important finding was related to the differences in benefits gained across experience levels. The system's structured scaffolding seems to benefit less experienced practitioners more which indicates that AI support can help mitigate some of the variance in decision quality that finding from differences in skill and experience. This potential deserves further empirical investigation in high-stakes professional contexts where consistency across practitioners is both quality demand and ongoing operational difficulty. This points to AI augmentation as a potential mechanism to raise the floor of collective judgement quality across diverse capability practitioner cohorts.

We conceptualised the AI augmented human support as a "third eye" in this study. This framing helped repositioning the AI-human divergences as a productive resource for reasoning-based judgement. The decision-making process then became more evidence centered, and reasoning based which supported individual judgments more visible and transparent. The consequential design implications are that developers, institutional actors and researchers should invest in designing interfaces and workflows that carefully surface and structure divergences as a mechanism to strengthen the quality and defensibility of human decisions.

There are risks associated with these affordances which need to be considered alongside benefits. The main attraction of using AI enabled processes is the cognitive ease which is also the factor most likely to result in automation bias. One of the key implications of this study is that well-structured and linguistically coherent AI generated outputs can create a false sense of accuracy and certainty which can lead to premature closure in judgement process without further inquiry. Our findings highlight that consistency and fluency in output should not be confused with quality and validity of underlying judgement which puts further importance on maintaining critical verification practices. AI systems that are based on historical data and existing institutional frameworks run the risk of reproducing the same patterns, biases, and inequities already present in those systems, rather than improving them. However, this does not mean AI should not be used in decision making. Instead, it emphasizes the necessity of robust governance systems that ensure human judgement and deliberation remain to be central in decision making, not just symbolically included.

In conclusion, this pilot suggests that the more important question is not whether AI can make expert work faster, but whether it can support better professional judgement quality. The findings indicate that, when implemented under carefully defined conditions, AI augmentation can enhance the consistency, accountability and equitable application of high-stakes decisions. However, these benefits depend on creating transparent, interrogable, and truly accountable systems which needs to leave the authority to human experts for decisions with AI serving as a tool for support rather than replacement.

Author contributions: ZE: conceptualization, data curation, formal analysis, investigation, methodology, project administration, validation, visualization, writing – original draft, writing – review & editing; **SL & RP:** supervision, validation, writing – review & editing; **MB:** investigation, resources, validation, writing – review & editing. All authors approved the final version of the article.

Funding: The authors received no financial support for the research and/or authorship of this article.

Acknowledgments: The authors would like to thank Nathan McCurley, Minnie Bumnanpol, Brodie Greguric, and Roman Fidyk from the Digital Technology Solutions AI team at the University of Newcastle for their technical expertise and support in the development of the AI-supported tool used in this study. Their contributions were instrumental in enabling the design and implementation of the system.

Ethics declaration: This study was approved by the Human Research Ethics Committee at the University of Newcastle on 24 January 2025 with approval number H-2024-0251. All participants provided informed consent prior to their involvement in the study. Candidate submissions used in the workshop were fully anonymized prior to analysis and discussion to protect candidate confidentiality and privacy.

AI statement: Generative AI was used solely to support language editing and stylistic polishing of the manuscript. The authors reviewed and revised all AI-assisted text and took full responsibility for the accuracy, interpretation, and final content of the article.

Declaration of interest: The authors declared no competing interest.

Data availability: Data generated or analyzed during this study are available from the authors on request.

REFERENCES

Australian Institute for Teaching and School Leadership. (2011). Australian professional standards for teachers. *AITSL*. https://www.aitsl.edu.au/teach/standards

Babic, M. J., White, R., Ersozlu, Z., Costigan, S., Kennedy, S., Guntoro, J., & Ledger, S. (2026). *A systematic review on teaching performance assessments impact on stakeholders* [Manuscript in preparation]. School of Education, The University of Newcastle.

Bansal, G., Nushi, B., Kamar, E., Lasecki, W., Weld, D., & Horvitz, E. (2021). Beyond accuracy: The role of mental models in human-AI team performance. In *Proceedings of the AAAI Conference on Human Factors in Computing Systems*.

Bloxham, S., & Boyd, P. (2012). Accountability in grading student work: Securing academic standards in a twenty-first century quality assurance context. *British Educational Research Journal, 38*(4), 615-634.

Bloxham, S., Hughes, C., & Adie, L. (2016). What's the point of moderation? A discussion of the purposes achieved through contemporary moderation practices. *Assessment & Evaluation in Higher Education, 41*(4), 638-653. https://doi.org/10.1080/02602938.2015.1039932

Boud, D., & Molloy, E. (2013). Rethinking models of feedback for learning: The challenge of design. *Assessment & Evaluation in Higher Education, 38*(6), 698-712. https://doi.org/10.1080/02602938.2012.691462

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77-101. https://doi.org/10.1191/1478088706qp063oa

Brooke, J. (1995). SUS: A quick and dirty usability scale. In P. W. Jordan, B. Thoms, B. A. Weerdmeester, & I. L. McClelland (Eds.), *Usability evaluation in industry* (pp. 189-194). Taylor & Francis.

Corbin, T., Bearman, M., Boud, D., & Dawson, P. (2025). The wicked problem of AI and assessment. *Assessment & Evaluation in Higher Education, 51*(4), 736-752. https://doi.org/10.1080/02602938.2025.2553340

Crompton, H., & Burke, D. (2023). Artificial intelligence in higher education: The state of the field. *International Journal of Educational Technology in Higher Education, 20*, Article 22. https://doi.org/10.1186/s41239-023-00392-8

Cummings, M. L. (2004). Automation bias in intelligent time critical decision support systems. In *Proceedings of the AIAA 1st Intelligent Systems Technical Conference*. https://doi.org/10.2514/6.2004-6313

Dai, W., Tsai, Y-S., Lin, J., Aldino, A., Jin, H., Li, T., Gasevic, D., & Chen, G. (2024). Assessing the proficiency of large language models in automatic feedback generation: An evaluation study. *Computers and Education: Artificial Intelligence, 7*, Article 100299. https://doi.org/10.1016/j.caeai.2024.100299

Fadel, C., Holmes, W., & Bialik, M. (2019). *Artificial intelligence in education: Promises and implications for teaching and learning*. Center for Curriculum Redesign.

Ferrara, E. (2024). Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *Science, 6*(1), Article 3. https://doi.org/10.3390/sci6010003

Flodén, J. (2025). Grading exams using large language models: A comparison between human and AI grading of exams in higher education using ChatGPT. *British Educational Research Journal, 51*, 201-224. https://doi.org/10.1002/berj.4069

Green, B., & Chen, Y. (2019). The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM Human Computer Interaction, 3*(CSCW), Article 50. https://doi.org/10.1145/3359152

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y., De Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors, 53*(5), 517-527. https://doi.org/10.1177/0018720811417254

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (task load index): Results of empirical and theoretical research. *Advances in Psychology, 52*, 139-183. https://doi.org/10.1016/S0166-4115(08)62386-9

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81-112. https://doi.org/10.3102/003465430298487

Henderson, M., Bearman, M., Chung, J., Fawns, T., Buckingham Shum, S., Matthews, K., & Heredia, J. (2025). Comparing generative AI and teacher feedback: Student perceptions of usefulness and trustworthiness. *Assessment & Evaluation in Higher Education*. https://doi.org/10.1080/02602938.2025.2502582

Holmes, W., Bialik, M., & Fadel, C. (2022a). *Artificial intelligence in education: Promise and implications for teaching and learning*. Center for Curriculum Redesign. https://doi.org/10.58863/20.500.12424/4276068

Holmes, W., Persson, J., Chounta, I.-A., Wasson, B., & Dimitrova, V. (2022b). *Artificial intelligence and education: A critical view through the lens of human rights, democracy and the rule of law*. Council of Europe.

Ilieva, G., Yankova, T., Ruseva, M., & Kabaivanov, S. (2025). A framework for generative AI-driven assessment in higher education. *Information, 16*(6), Article 472. https://doi.org/10.3390/info16060472

Ioannidis J. P. A. (2005). Why most published research findings are false. *PLoS Medicine, 19*(8), Article e1004085. https://doi.org/10.1371/journal.pmed.0020124

Joughin, G. (2008). Assessment, learning and judgement in higher education: A critical review. In G. Joughin (Ed.), *Assessment, learning and judgement in higher education* (pp. 1-15). Springer. https://doi.org/10.1007/978-1-4020-8905-3_2

Kellogg, R. T. (2008). Training writing skills: A cognitive developmental perspective. *Journal of Writing Research, 1*(1), 1-26. https://doi.org/10.17239/jowr-2008.01.01.1

Klenowski, V., & Adie, L. (2009). Moderation as judgement practice. *Assessment & Evaluation in Higher Education, 34*(1), 1-11.

Lee, J. D., & See, K. A. (2004). Trust in automation. *Human Factors, 46*(1), 50-80. https://doi.org/10.1518/hfes.46.1.50.30392

Li, J., Jangamreddy, N. K., Hisamoto, R., Bhansali, R., Dyda, A., Zaphir, L., & Glencross, M. (2024). AI-assisted marking: Functionality and limitations of ChatGPT in written assessment evaluation. *Australasian Journal of Educational Technology, 40*(4), 56-72. https://doi.org/10.14742/ajet.9463

Liao, V., & Varshney, K. (2021). *Human-centered explainable AI (XAI): From algorithms to user experiences*. arXiv. https://doi.org/10.48550/arXiv.2110.10790

Luckin, R., Holmes, W., Griffiths, M., & Forcier, L. B. (2016). *Intelligence unleashed: An argument for AI in education*. Pearson.

Mascadri, J., Spina, N., Spooner-Lane, R., & Alonzo, D. (2023). Exploring consistency and variation in assessor judgements in teacher performance assessment. *Assessment & Evaluation in Higher Education, 48*(7), 1085-1100.

Miller, T. (2017). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267*, 1-38. https://doi.org/10.1016/j.artint.2018.07.007

Moorhouse, B., Yeo, M., & Wan, Y. (2023). Generative AI tools and assessment: Guidelines of the world's top-ranking universities. *Computers and Education Open, 5*, Article 100151. https://doi.org/10.1016/j.caeo.2023.100151

Mosier, K. L., Skitka, L. J., Heers, S., & Burdick, M. (1998). Automation bias. *International Journal of Human-Computer Studies, 52*(4), 701-717.

Oberlader, V., Quinten, L., Schmidt, A. F., & Banse, R. (2025). How can I reduce bias in my work? Discussing debiasing strategies for forensic psychological assessments. *Professional Psychology: Research and Practice, 56*(3), 211-221. https://doi.org/10.1037/pro0000615

Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist, 38*(1), 1-4. https://doi.org/10.1207/S15326985EP3801_1

Parasuraman, R., & Riley, V. (1997). Humans and automation. *Human Factors, 39*(2), 230-253. https://doi.org/10.1518/001872097778543886

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics – Part A, 30*(3), 286-297. https://doi.org/10.1109/3468.844354

Raji, I., Smart, A., White, R., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 33-44). ACM. https://doi.org/10.1145/3351095.3372873

Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences, 20*(9), 676-688. https://doi.org/10.1016/j.tics.2016.07.002

Sadler, D. R. (2013). Assuring academic achievement standards. *Assessment in Education, 20*(1), 5-16. https://doi.org/10.1080/0969594X.2012.714742

Sinha, S., Goel, S., Kumaraguru, P., Geiping, J., Bethge, M., & Prabhu, A. (2025). *Can language models falsify? Evaluating algorithmic reasoning with counterexample creation*. arXiv. https://doi.org/10.48550/arXiv.2502.19414

Skitka, L. J., Mosier, K. L., & Burdick, M. (1999). Does automation bias decision-making? *International Journal of Human-Computer Studies, 51*(5), 991-1006. https://doi.org/10.1006/ijhc.1999.0252

Stacey, M., Talbot, D., & Buchanan, J. (2020). Teacher performance assessments and the making of professional standards. *Teaching and Teacher Education, 95*, Article 103145.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*(2), 257-285. https://doi.org/10.1207/s15516709cog1202_4

Sweller, J. (2011). Cognitive load theory. In J. P. Mestre, & B. H. Ross (Eds.), *The psychology of learning and motivation: Cognition in education* (pp. 37-76). Elsevier Academic Press. https://doi.org/10.1016/B978-0-12-387691-1.00002-8

Swiecki, Z., Khosravi, H., Chen, G., Martinez-Maldanado, R., Lodge, J., Milligan, S., Selwyn, N., & Gasevic, D. (2022). Assessment in the age of artificial intelligence. *Computers and Education: Artificial Intelligence, 3*, Article 100075. https://doi.org/10.1016/j.caeai.2022.100075

Usher, M. (2025). Generative AI vs. instructor vs. peer assessments: A comparison of grading and feedback in higher education. *Assessment & Evaluation in Higher Education, 50*(6), 912-927. https://doi.org/10.1080/02602938.2025.2487495

Weston, H., Mooney, A., Thomas, M. K. E., & Iannucci, C. (2026). Graduate teachers' and the teacher performance assessment: The case for architectures of judgement. *Assessment & Evaluation in Higher Education*. https://doi.org/10.1080/02602938.2026.2658631

Williamson, B., & Eynon, R. (2020). Historical threads, missing links, and future directions in AI in education. *Learning, Media and Technology, 45*(3), 223-235. https://doi.org/10.1080/17439884.2020.1798995

Wyatt-Smith, C., Adie, L., & Harris, L. (2024). Supporting teacher judgement and decision-making: Using focused analysis to help teachers see students, learning, and quality in assessment data. *British Educational Research Journal, 50*(3), 1420-1448. https://doi.org/10.1002/berj.3984

Zacharis, G., & Papadakis, S. (2025). Can AI grade like a human? Validity, reliability, and fairness in university coursework assessment. *Educational Process: International Journal, 19*, Article e2025591. https://doi.org/10.22521/edupij.2025.19.591

Zawacki-Richter, O., Marin, V. I., Bond, M & Gouverneur, F. (2019). Systematic review of AI applications in higher education. *International Journal of Educational Technology in Higher Education, 16*, Article 39. https://doi.org/10.1186/s41239-019-0171-0