

Human Factors in AI-Assisted Grading: A Mixed-Methods Study of Efficiency, Workload, and Interaction Patterns

Jamlech Iram N. Gojo Cruz^{1,2,*}, Maria Aura Teodora C. Matias³

¹*Institute of Computer Science, University of the Philippines Los Baños, Philippines*

²*Artificial Intelligence Program, University of the Philippines Diliman, Philippines*

³*Industrial Engineering & Operations Research, University of the Philippines Diliman, Philippines*

Abstract

This study examines human factors in AI-assisted grading through a mixed-methods design, comparing ChatGPT-assisted and manual grading among ten computer science faculty. A within-subjects experiment measured grading efficiency and perceived workload, while contextual inquiry and post-task interviews examined interaction patterns, trust, and behavioral responses. A paired t-test found no significant difference in grading time ($t(9) = 0.317$, $p = .379$, $d = 0.100$). A one-tailed Wilcoxon signed-rank test found a significant and large reduction in perceived workload favoring ChatGPT ($W = 21.0$, $p = .016$, $r = .696$). Three grader personas emerged: AI-first, AI-as-reference, and human-first which reflects a trust-driven spectrum of automation reliance and educator agency. Faculty awarded higher scores on average ($d = 0.708$) and overrode ChatGPT's zero scores in 58.3% of cases; this demonstrates active corrective judgment rather than passive deference. Anthropomorphism and perceived authority shift were observed qualitatively. These findings establish that evaluating AI grading tools on technical accuracy alone is insufficient. Persona-adaptive interfaces, transparency mechanisms, and governance structures that reinforce educator agency are essential for responsible human-AI interaction in assessment contexts.

Keywords

AI-assisted grading, anthropomorphism, automation bias, ChatGPT, human-AI interaction, trust

1. Introduction

Grading is among the most time-intensive tasks educators perform, and tools like ChatGPT offer a potentially transformative form of support. Despite growing adoption of AI-assisted grading in educational settings, little empirical work has directly compared its performance against manual grading under controlled conditions. Most existing studies evaluate AI grading by technical accuracy alone: how closely AI scores match a gold standard. Whether they actually save time or reduce the cognitive burden on faculty has not been directly measured. Understanding these practical implications is a foundational question for educators and institutional decision-makers.

A related gap concerns how faculty actually behave when working alongside AI grading tools. Even if an AI tool is accurate and fast, its practical value depends on whether faculty exercise sufficient independent judgment in using it. Prior human-computer interaction research identifies automation bias and anthropomorphism as constructs that can distort how users engage with AI systems[1, 2], yet neither has been examined, or has only been limitedly examined, in the context of educational assessment. The extent to which AI grading tools support or undermine educator agency, the capacity to exercise professional judgment and retain meaningful control over assessment decisions, remains an open question.

This study aims to provide empirical insights and design implications on human-AI interaction in grading tasks by comparing ChatGPT-assisted grading against manual grading. Specifically, the paper aims to answer the following research questions:

HAI-Agency: Workshop on Orchestrating Human and AI Agency for Proactive and Reflective Learning, co-located with the 27th International Conference on Artificial Intelligence in Education (AIED 2026), Seoul, Korea, June 27 - July 3, 2026.

*Corresponding author.

✉ jngojocruz@up.edu.ph (J.I.N. Gojo Cruz); acmatias@up.edu.ph (M. A. T. C. Matias)

ORCID 0009-0001-8164-8778 (J.I.N. Gojo Cruz)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

- RQ1: How does ChatGPT-assisted grading compare to manual grading in terms of efficiency (time taken)?
- RQ2: How does ChatGPT use influence perceived workload during grading?
- RQ3: What are faculty perceptions of ChatGPT-assisted grading in terms of feedback quality, trust, and willingness to adopt?
- RQ4: What interaction patterns emerged, and how do they inform human-centered design of AI grading tools?

RQ1 and RQ2 address the experimental component; RQ3 and RQ4 address the behavioral component. Together, they provide a complementary picture of AI-assisted grading as both a practical tool and a human-AI interaction context with direct implications for educator agency.

2. Background

2.1. AI-Assisted Grading in Education

AI tools have demonstrated considerable promise across nearly all phases of educational measurement, from item creation to automating scoring of student submissions [3, 4]. In computer science education specifically, ChatGPT has shown utility for evaluating code and providing feedback [5], though it struggles with higher-order programming tasks and requires substantial teacher oversight [6]. Therefore, a human-in-the-loop approach remains essential: Tomić et al. [7] demonstrate this in collaborative grading contexts; the U.S. Department of Education [8] recommends teacher oversight as a governance principle; and Jeon and Lee [9] argue that LLMs function best as complements to, rather than replacements for, human educators. These are all to ensure quality, fairness, and meaningful learning outcomes.

Studies consistently note that AI grading can reduce time and workload [10, 11] and offer immediate and detailed feedback, even at scale [12, 13, 14, 15]. However, these studies share a common limitation: they evaluate AI tools almost exclusively on technical accuracy, leaving unexamined how faculty actually interact with, verify, and respond to AI output in practice, and how trust, workload, and behavioral adaptation shape the real-world utility of these tools.

2.2. Trust, Automation Bias, and Educator Agency

Trust is a central construct in human-automation interaction. Lee and See [16] define trust in automation as the attitude that an agent will help achieve an individual's goals under uncertainty. They distinguish between appropriate reliance, where trust is calibrated to the system's reliability, and miscalibrated trust, which manifests as over-reliance or under-reliance. A key consequence of over-reliance is automation bias: the tendency to favor automated recommendations over independent judgment, even when automation produces errors. Automation bias is especially likely when AI output is presented with high confidence, when users are under time pressure, and when users have limited experience with the system's failure modes. These conditions are commonly present in grading contexts [1, 17].

3. Method

This study employed a mixed-methods design combining a within-subjects experiment with a concurrent behavioral inquiry. Figure 1 provides an overview of the study structure.

3.1. Participants

Ten computer science faculty (8 male, 2 female; age 18–44; 1–15 years teaching experience; positions ranging from Instructor to Assistant Professor) participated in the study. Ethical approval was obtained under a faculty-level review protocol. All participants provided written informed consent prior to participation and were assured of anonymity.

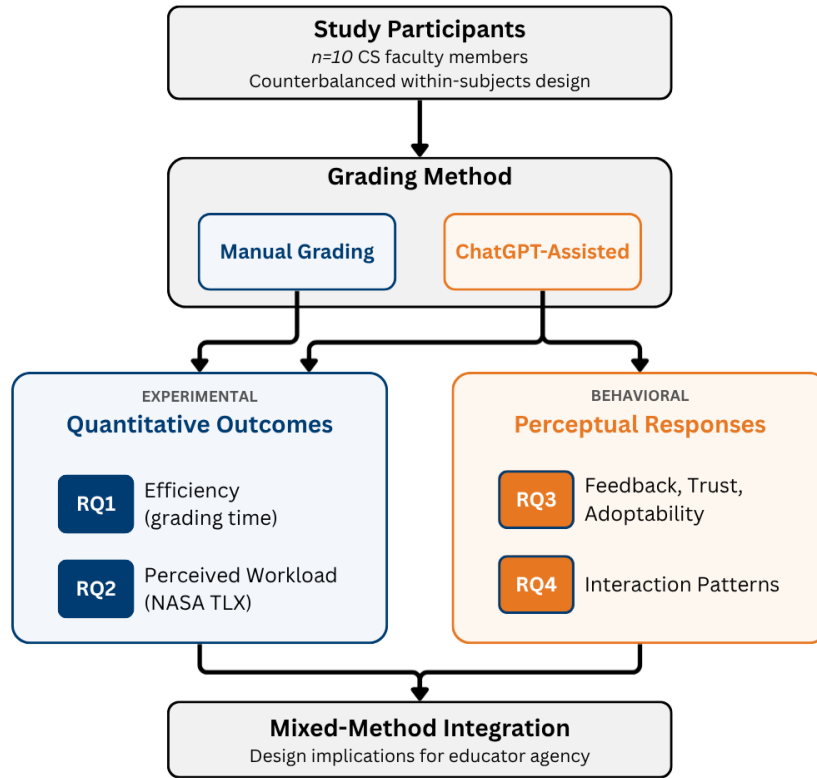


Figure 1: Study design overview.

3.2. Experimental Design

A within-subjects design was employed with grading method as the independent variable: manual grading versus ChatGPT-assisted grading and a counterbalanced order was employed to mitigate the carryover effects. The grading task involved evaluating ten student introductory Python programming submissions against a rubric (10 base points + 1 bonus).

In the ChatGPT-assisted condition, AI evaluations were pre-generated by the researcher prior to the study session and presented to participants as static text documents alongside the student submissions. Each evaluation was generated using a structured prompt that included: (1) the full student code submission, (2) the grading rubric with point allocations, (3) an explicit instruction to award partial credit where appropriate, and (4) contextual framing that submissions were from beginner-level students. The full prompt is provided in Appendix B. This design choice served two purposes: it controlled for variability in participant prompting behavior, and it ensured that the perceived workload measure reflected the cognitive cost of reviewing and adjudicating AI output specifically, rather than the added cost of navigating a conversational interface. The ChatGPT model used was o4-mini-high for the ChatGPT-assisted grading condition.

Two dependent variables were measured: (1) efficiency, operationalized as task completion time in minutes, and (2) perceived workload, measured using the Raw NASA Task Load Index (Raw TLX) [18], which yields an unweighted average across six dimensions (mental demand, physical demand, temporal demand, performance, effort, frustration) rated 0–7. Based on AI’s capacity to automate and assist in coding tasks, we hypothesize that manual grading will show higher time taken and higher perceived workload than ChatGPT-assisted grading. This hypothesis is the basis of the statistical analyses.

3.3. Qualitative Component

Concurrent contextual inquiry [19] was conducted during the experiment, followed by semi-structured post-task interviews. Interviews explored feedback quality, trust, willingness to adopt, and design

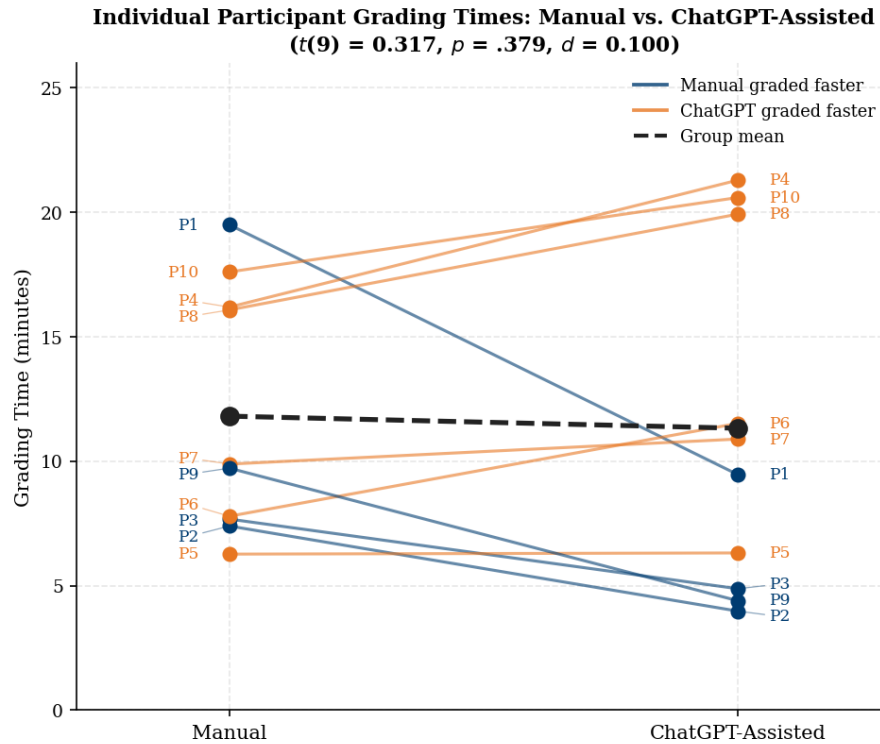


Figure 2: Individual participant grading times by condition.

preferences. These dimensions were not selected ad hoc: feedback quality maps to perceived usefulness and trust maps to Lee and See's [16] calibration framework, while willingness to adopt captures behavioral intention, the outcome variable in the Technology Acceptance Model [20]. Design preferences were included as a forward-looking dimension to generate actionable implications. Audio recordings were transcribed, translated (English/Filipino), and analyzed using thematic analysis [21], with ChatGPT used as a supplementary analytical aid and all patterns verified against raw transcripts.

4. Results

4.1. Efficiency

A Shapiro-Wilk test confirmed normality of within-subject time differences ($SW = 0.927$, $p = .419$). A one-tailed paired t -test found no significant difference in grading time between conditions ($t(9) = 0.317$, $p = .379$, $d = 0.100$, 95% CI: $[-2.98, 3.95]$ min). This result should be treated as inconclusive given the study's low statistical power (approximately 8.8%) rather than as evidence of equivalence. Mean grading times were 11.80 min ($SD = 4.96$) for manual and 11.32 min ($SD = 6.92$) for ChatGPT-assisted. Individual-level data revealed wide variability: six participants graded faster manually while four graded faster with ChatGPT. Figure 2 shows this individual-level grading time, and it illustrates the wide dispersion of slope directions which indicate that ChatGPT did not reduce grading time uniformly.

4.2. Perceived Workload

A Wilcoxon signed-rank test (non-normality confirmed: $SW = 0.758$, $p = .005$) found a statistically significant and large reduction in perceived workload favoring ChatGPT-assisted grading ($W = 21.0$, $p = .016$, $r = .696$). Manual grading yielded higher overall Raw TLX scores ($M = 4.32$, $SD = 0.95$) than ChatGPT-assisted grading ($M = 3.83$, $SD = 0.73$). All six non-tied participants reported lower workload under ChatGPT assistance. Temporal demand and effort showed the largest reductions (both $r = .696$, $p =$

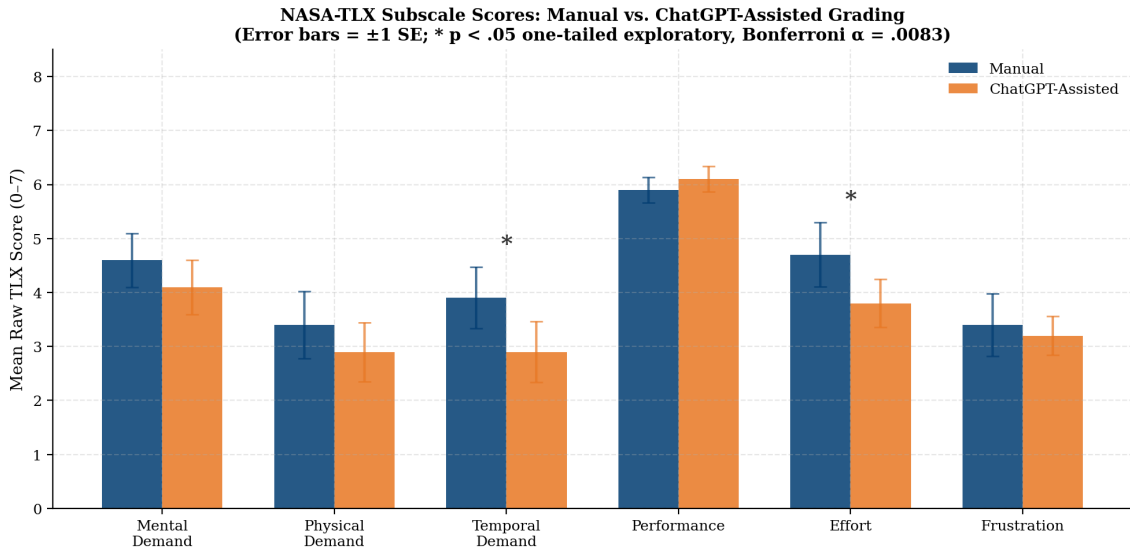


Figure 3: Mean Raw NASA-TLX subscale scores by grading condition.

Table 1

NASA-TLX Subscale Comparison Between Manual and ChatGPT-Assisted Grading.

Subscale	Manual <i>M (SD)</i>	ChatGPT <i>M (SD)</i>	Mean Diff	<i>n</i>	<i>W</i>	<i>p</i>	<i>r</i>
Mental Demand	4.60 (1.58)	4.10 (1.60)	+0.50	5	13.0	.125	.469
Physical Demand	3.40 (1.96)	2.90 (1.73)	+0.50	5	10.0	.344	.213
Temporal Demand	3.90 (1.79)	2.90 (1.79)	+1.00	6	21.0	.016*	.696
Performance	5.90 (0.74)	6.10 (0.74)	−0.20	6	7.0	.891	.232
Effort	4.70 (1.89)	3.80 (1.40)	+0.90	6	21.0	.016*	.696
Frustration	3.40 (1.84)	3.20 (1.14)	+0.20	7	15.0	.469	.053
Overall	4.32 (0.95)	3.83 (0.73)	+0.48	6	21.0	.016*	.696

Note. One-tailed Wilcoxon signed-rank test (alternative = “greater”: Manual > ChatGPT), exact distribution, $N = 10$. n = number of non-zero paired differences; W = sum of positive ranks; $r = |z|/\sqrt{N}$, where z is from the normal approximation. Subscale analyses are exploratory; Bonferroni-corrected threshold $\alpha = .0083$ — no subscale survives correction. For Performance, the reversed direction (ChatGPT > Manual) explains the high one-tailed p -value. Effect size conventions: small ($r = .10$), medium ($r = .30$), large ($r = .50$). * $p < .05$ (one-tailed, exact).

.016), as shown in Figure 3. This is consistent with cognitive offloading, that when AI pre-structures the grading output, faculty transition from generating assessments to evaluating them, reducing intrinsic cognitive load without necessarily reducing total time.

Notably, for temporal demand, effort, and the overall workload measure, all non-tied participants reported reductions under ChatGPT assistance, the maximum directional consistency the test can detect at this sample size (see Table 1).

4.3. Perceptions of ChatGPT-Assisted Grading

4.3.1. Feedback Quality

The dominant perception was that ChatGPT graded too harshly. Despite explicit prompting to award partial credit and account for beginner-level context, ChatGPT frequently assigned zeros. Participants noted: “It is too harsh... the AI lacks humanity” (P5); “If it does not work, it gives a zero... no partial credit” (P3). Positive perceptions centered on consistency and speed, though some found feedback

Table 2

Observed Interaction Patterns Among Participants During ChatGPT-Assisted Grading.

Persona	Description	Trust Characteristics
AI-first	Participants displayed high trust in ChatGPT and readily adopted its evaluations with minimal review or independent rechecking.	Self-reported trust $\geq 90\%$. Minimal verification.
AI-as-reference	Participants reviewed ChatGPT's evaluations first but verified them manually, using ChatGPT primarily as a validation tool and starting point.	Moderate trust ($\approx 60\text{--}75\%$). Treats AI output as a starting point rather than a conclusion.
Human-first	Participants conducted their own grading independently and only consulted ChatGPT as a secondary reference to check for overlooked errors or to obtain a second opinion.	Lower trust ($\leq 60\%$ or unstated). Preserves independent judgment as the primary basis for final grades.

verbose and observed false positives.

4.3.2. Trust

Self-reported trust ranged from 40% to 90% and was calibrated to how closely ChatGPT's reasoning aligned with participants' own judgment. Higher-trust participants accepted AI evaluations with minimal review: *"I am confident, especially when its logic aligns with mine,"* (P2); *"Mostly, I rely on the grades ChatGPT gave... I then just verify them"* (P1). Lower-trust participants maintained independent scrutiny: *"I am not confident... as is,"* (P8); *"I still edit the score and feedback"* (P6). Participants were more receptive to ChatGPT's qualitative feedback than its numeric grades.

4.3.3. Adoptability

No participant rejected future adoption outright. Willingness was conditional on improved accuracy, better partial credit handling, and growing familiarity. Participants saw stronger potential for ChatGPT in grading larger or more complex assignments.

4.4. Interaction Patterns and Grader Personas

Three distinct interaction personas emerged from contextual inquiry and post-task interviews, differing primarily in trust level and whether AI output was consulted before or after forming an independent judgment. Table 2 summarizes these patterns.

Beyond the three personas, contextual inquiry revealed three additional behavioral patterns. First, anthropomorphism: several participants framed ChatGPT as a collaborator, for instance: *"It feels like I have a college classmate as my partner"*. Second, perceived authority shift: *"Instead of me having the authority, it feels like ChatGPT has authority... I changed my behavior because of its suggestions"* (P10). Third, a corrective impulse: several participants actively audited ChatGPT's judgments against their own pedagogical standards rather than passively accepting AI output, and in some cases explicitly articulated what they would want the AI to do differently if given the opportunity. Faculty were not merely consumers of AI output but active output auditors who intervened where AI judgment fell short. The first two patterns represent meaningful threats to educator agency, while the third represents an underlying resource. Interface design and governance must address both dimensions.

4.5. Grading Quality: Supplementary Analysis

Faculty awarded higher total scores on average ($M = 8.52$) than ChatGPT suggested ($M = 7.60$, $t(9) = 2.240$, $p = .052$, $d = 0.708$). Viewed at the aggregate level, this difference did not reach statistical significance, suggesting that ChatGPT's overall score trajectory was broadly aligned with faculty judgment. Spearman's rank correlation between ChatGPT-suggested and faculty-assigned total scores

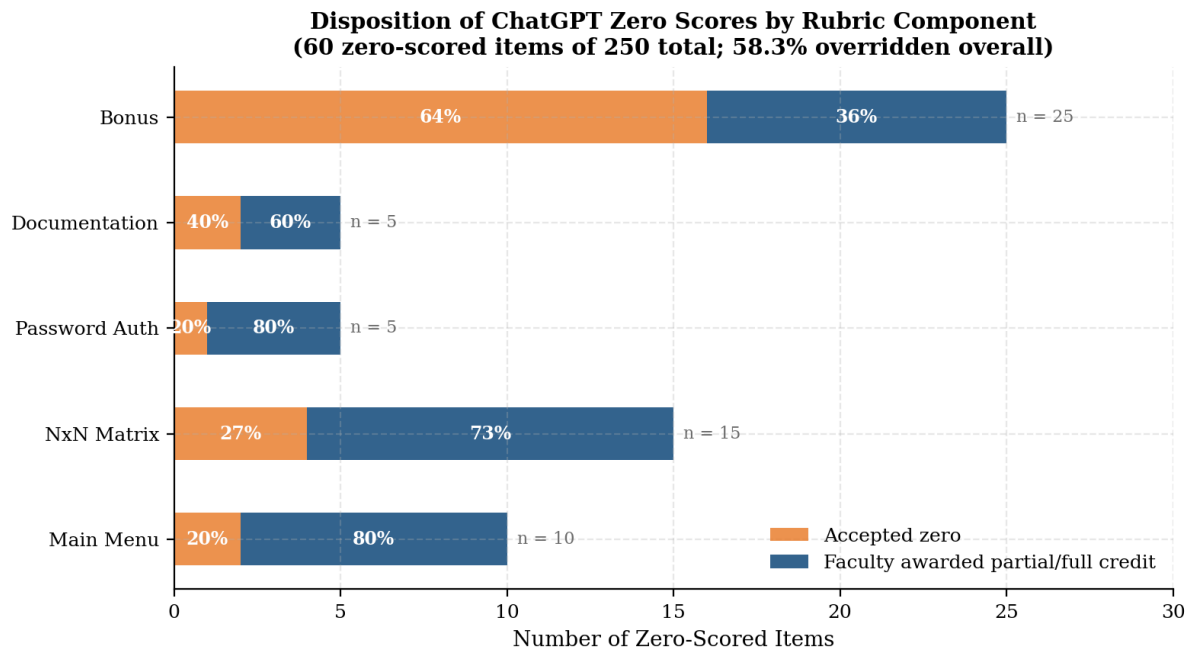


Figure 4: Disposition of ChatGPT zero scores by rubric component.

was $\rho = .859$, indicating strong ordinal agreement in how submissions were ranked even where absolute scores differed.

At the item level, however, a more granular picture emerges: across 250 rubric items, ChatGPT assigned 60 zeros; faculty accepted only 25 (41.7%) and overrode the remaining 35 (58.3%), applying partial credit where ChatGPT gave none. Override rates, as shown in Figure 4 were highest for core rubric components (Main Menu: 80%; Password Auth: 80%; NxN Matrix: 73.3%). Aside from the active corrective judgment, these results suggest that faculty as a group maintained educator agency even when AI output was influential.

5. Discussion

5.1. Efficiency, Workload, and Cognitive Offloading

The absence of an efficiency advantage and the significant workload reduction present a notable contrast. The individual-level grading time data suggest that trust mediated any potential time savings: participants with lower trust spent additional time verifying ChatGPT's output, counteracting efficiency gains that may exist for high-trust users, meaning efficiency benefits may be persona-dependent rather than absent. The workload reduction finding is more robust and theoretically significant.

Although ChatGPT did not save time, faculty perceived the task as substantially less demanding particularly in temporal demand and effort. This is consistent with cognitive offloading theory [22]: an external resource reduces intrinsic cognitive demands even when total time is unchanged. When the grading output is already given, faculty shift from manually scrutinizing the code line-by-line to focusing on what AI points them to, and in return reducing the cognitive cost of the task. This workload benefit has standalone practical value: faculty who perceive grading as less demanding are more likely to sustain engagement and provide higher-quality feedback over time.

5.2. Feedback Quality, Fairness, and Scoring

ChatGPT's systematic harshness (e.g., defaulting to zero scores without consideration of other factors) illustrates a known challenge in automated assessment, that the algorithmic consistency does not

guarantee pedagogical fairness when the algorithm lacks contextual grounding [23]. The path to a final grade matters as well as the grade itself: the majority of the faculty acknowledge incremental progress and consider the capabilities of their students beyond their programming output. Faculty consistently recognized this distinction: they awarded partial credit on 58.3% of items ChatGPT scored as zero, and the high inter-rater reliability (ICC = 0.977 overall) confirms that this corrective behavior was consistent across graders. This demonstrates the kind of contextual, holistic judgment that automated systems cannot replicate. Moreover, the faculty's active override behavior shows that professional judgment was maintained even when AI output was influential.

5.3. Automation Bias, Anthropomorphism, and Agency

The results present a nuanced picture of automation bias. At the individual level, contextual inquiry revealed susceptibility. Some participants revised scores after viewing ChatGPT's output without independently re-examining student code, and some participants explicitly described a perceived authority shift. At the aggregate level, however, faculty awarded higher scores and overrode AI zeros in the majority of cases. This indicates that automation bias was felt but largely resisted. The AI-first graders (n=2) represent the exception: their minimal verification behavior at self-reported trust $\geq 90\%$ is consistent with commission-type automation bias [1], and constitutes a governance concern in higher-stakes grading contexts even when the broader group exercises appropriate skepticism.

Anthropomorphism amplified emotional reactivity in ways that interface design must address. When faculty framed ChatGPT as a collaborator, disagreements with its output were experienced as interpersonal conflicts. This is consistent with prior work on anthropomorphism, where people tend to attribute human-like qualities to agents whose outputs appear intentional and contextually responsive [2], both of which characterize conversational LLMs like ChatGPT. The emotional consequences were observable: agreement with ChatGPT produced reassurance, while disagreement produced frustration directed at the AI rather than the student's code. One participant noted: *"I was more irritated with ChatGPT, because I was arguing with it, not with the code."* This displacement represents a concrete mechanism by which anthropomorphism can distort judgment, and adds an affective dimension to the agency dynamics that governance frameworks should account for.

5.4. Grader Personas and Trust

The three personas map onto the trust calibration framework discussed in the literature [16]. AI-first graders sit closest to over-reliance; human-first graders closest to calibrated trust. The distribution (i.e., two AI-first, four AI-as-reference, four human-first) suggests most faculty approached the tool cautiously and thus preserved agency. Trust emerged as the central determinant of persona assignment shaped by initial impressions formed during the session.

5.5. Limitations

There are limitations that bound the interpretation of these findings. First, the sample of ten faculty members yields low statistical power at approximately 8.8% for the efficiency comparison. This means that the null RQ1 result is inconclusive rather than definitive, and effect sizes should be treated as preliminary estimates pending replication with larger samples. Second, the task was restricted to beginner-level Python exercises involving iterative statements; grading more complex or open-ended assignments may produce different efficiency and workload profiles, and ChatGPT's limitations may be more pronounced or different in nature. Third, the qualitative analysis followed an interpretive approach without independent inter-rater coding; all patterns were verified against raw transcripts to ensure grounding in the data. Finally, none of the participants had prior experience using ChatGPT for grading, meaning the trust levels and interaction patterns observed reflect initial impressions rather than calibrated, experience-informed reliance. Longitudinal studies tracking how grader personas and trust evolve with sustained tool use would significantly strengthen the findings reported here.

6. Design Implications

6.1. Persona-Adaptive Interfaces

To preserve educator agency across trust levels, AI grading interfaces should adapt to individual grader personas. The following recommendations are grounded in participant responses from the contextual inquiry and post-task interviews. For AI-first graders, interfaces should include mandatory verification prompts and confidence indicators that surface the AI's reasoning and highlight cases where independent review is warranted. For AI-as-reference graders, side-by-side layouts that display suggested scores alongside editable fields support efficient verification without displacing judgment. For human-first graders, on-demand AI assistance (i.e., generated only upon request or used post-grading as an audit) respects their preference to ground decisions in independent evaluation first. Over time, systems could learn from individual grading behavior to align suggestions with a teacher's own style and contextual standards.

Beyond persona adaptation, interfaces should formalize the underlying corrective impulse observed in this study: providing teachers with lightweight mechanisms to flag AI judgments they disagree with and to communicate rubric interpretation preferences back to the system. This closes the human-AI feedback loop and positions teachers not just as auditors of AI output but as active shapers of AI behavior over time.

6.2. AI Governance Guidelines

Given the evidence of automation bias susceptibility and perceived authority shift, governance frameworks must treat these risks as baseline concerns rather than edge cases. Systems should make explicit how scores and feedback were generated. Confidence indicators and rationale panels can help teachers interrogate AI reasoning rather than simply accept results [24, 25]. However, such mechanisms must be designed carefully, as explanations can paradoxically increase over-reliance when users trust an output simply because an explanation is present [26]. Accountability must remain clearly with the teacher: AI tools must be positioned as decision-support, not decision-makers. Lastly, privacy must be safeguarded when systems learn from student submissions or teacher behavior, particularly if personalization features are introduced.

7. Conclusion

This study examined ChatGPT-assisted grading through two complementary lenses: a controlled experiment measuring efficiency and perceived workload, and contextual inquiry examining behavioral interaction patterns. Our key findings are:

1. ChatGPT did not significantly reduce grading time but did significantly reduce perceived workload, consistent with cognitive offloading.
2. Faculty were identified as AI-first, AI-as-reference, or human-first based on trust-driven differences in how they orchestrated their own agency relative to the AI.
3. Even though automation bias susceptibility was observed in some individuals, faculty as a group exercised professional autonomy and intervened where they thought AI judgment fell short.
4. Anthropomorphism and perceived authority shift were observed qualitatively, representing affective risks to educator agency that interface design must address.

These findings contribute empirical grounding to questions of human-AI agency in education. The grader personas offer a behavioral model for understanding how educators navigate the boundary between AI assistance and professional autonomy. Designing AI-assisted grading tools that preserve educator agency requires persona-adaptive interfaces, transparent reasoning, and governance structures that keep teachers in control. This is not merely a design preference but a psychological and ethical necessity. As AI becomes more deeply embedded in academic practice, understanding how it shapes human judgment will be as important as evaluating what it produces.

Declaration on Generative AI

During the preparation of this work, the author(s) used Claude Sonnet 4.6 (Anthropic) in order to improve the readability and language of portions of the text. After using this tool, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

- [1] R. Parasuraman, D. H. Manzey, Complacency and bias in human use of automation: An attentional integration, *Human Factors* 52 (2010) 381–410. URL: <https://doi.org/10.1177/0018720810376055>. doi:10.1177/0018720810376055.
- [2] N. Epley, A. Waytz, J. T. Cacioppo, On seeing the human: A three-factor theory of anthropomorphism, *Psychological Review* 114 (2007) 864–886. URL: <https://doi.org/10.1037/0033-295X.114.4.864>. doi:10.1037/0033-295X.114.4.864.
- [3] V. J. Owan, K. B. Abang, D. O. Idika, E. O. Etta, B. A. Bassey, Exploring the potential of artificial intelligence tools in educational measurement and assessment, *Eurasia Journal of Mathematics, Science and Technology Education* 19 (2023). URL: <https://doi.org/10.29333/ejmste/13428>. doi:10.29333/ejmste/13428.
- [4] M. van Haastrecht, L. de Groot, M. Jongbloed-Pereboom, F. Buytenhuijs, J. Kruis, Artificial intelligence for educational measurement: Where is the value for education?, *Frontiers in Education* Volume 10 - 2025 (2025). URL: <https://www.frontiersin.org/journals/education/articles/10.3389/feduc.2025.1677255>. doi:10.3389/feduc.2025.1677255.
- [5] M. B. Garcia, Teaching and learning computer programming using chatgpt: A rapid review of literature amid the rise of generative ai technologies, *Education and Information Technologies* 30 (2025) 16721–16745. URL: <https://doi.org/10.1007/s10639-025-13452-5>. doi:10.1007/s10639-025-13452-5.
- [6] Z. He, T. Nguyen, T. Miari, M. Aliasgari, S. Rafatirad, H. Sayadi, The ai companion in education: Analyzing the pedagogical potential of chatgpt in computer science and engineering, in: 2024 IEEE Global Engineering Education Conference (EDUCON), 2024, pp. 1–10. URL: <https://doi.org/10.1109/educon60312.2024.10578820>. doi:10.1109/educon60312.2024.10578820.
- [7] B. B. Tomić, A. D. Kijevčanin, Z. V. Ševarec, J. M. Jovanović, An ai-based approach for grading students' collaboration, *IEEE Transactions on Learning Technologies* 16 (2023) 292–305. URL: <https://doi.org/10.1109/TLT.2022.3225432>. doi:10.1109/TLT.2022.3225432.
- [8] M. A. Cardona, R. J. Rodriguez, K. Ishmael, Artificial intelligence and the future of teaching and learning: Insights and recommendations, 2023. URL: <https://www.ed.gov/sites/ed/files/documents/ai-report/ai-report.pdf>.
- [9] J. Jeon, S. Lee, Large language models in education: A focus on the complementary relationship between human teachers and chatgpt, *Education and Information Technologies* 28 (2023) 15873–15892. URL: <https://doi.org/10.1007/s10639-023-11834-1>. doi:10.1007/s10639-023-11834-1.
- [10] D. Deepshikha, A comprehensive review of ai-powered grading and tailored feedback in universities, *Discover Artificial Intelligence* 5 (2025) 251. URL: <https://doi.org/10.1007/s44163-025-00517-0>. doi:10.1007/s44163-025-00517-0.
- [11] Y. Li, M. Raković, N. Srivastava, X. Li, Q. Guan, D. Gašević, G. Chen, Can ai support human grading? examining machine attention and confidence in short answer scoring, *Computers & Education* 228 (2025) 105244. URL: <https://www.sciencedirect.com/science/article/pii/S0360131525000120>. doi:<https://doi.org/10.1016/j.compedu.2025.105244>.
- [12] W. Dai, J. Lin, H. Jin, T. Li, Y.-S. Tsai, D. Gašević, G. Chen, Can large language models provide feedback to students? a case study on chatgpt, in: 2023 IEEE international conference on advanced learning technologies (ICALT), IEEE, 2023, pp. 323–325. doi:10.1109/ICALT58122.2023.00100.
- [13] M. Jukiewicz, The future of grading programming assignments in education: The role of chatgpt

- in automating the assessment and feedback process, *Thinking Skills and Creativity* 52 (2024) 101522. URL: <https://www.sciencedirect.com/science/article/pii/S1871187124000609>. doi:<https://doi.org/10.1016/j.tsc.2024.101522>.
- [14] A. McGowan, N. Anderson, C. Smith, The use of chatgpt to generate summative feedback in programming assessments that is consistent, prompt, without bias and scalable, in: *Proceedings of the Cognitive Models and Artificial Intelligence Conference*, 2024, pp. 39–43. doi:10.1145/3660853.3660863.
- [15] G. Nagakalyani, S. Chaudhary, V. Apte, G. Ramakrishnan, Ta buddy: Ai-assisted grading tool for introductory programming assignments, in: *Proceedings of the 56th ACM Technical Symposium on Computer Science Education V. 2 (SIGSETS 2025)*, Association for Computing Machinery, 2025, p. 1736. URL: <https://doi.org/10.1145/3641555.3705051>. doi:10.1145/3641555.3705051.
- [16] J. D. Lee, K. A. See, Trust in automation: Designing for appropriate reliance, *Human Factors* 46 (2004) 50–80. URL: https://journals.sagepub.com/doi/10.1518/hfes.46.1.50_30392. doi:10.1518/hfes.46.1.50_30392.
- [17] A. Klingbeil, C. Grützner, P. Schreck, Trust and reliance on ai — an experimental study on the extent and costs of overreliance on ai, *Computers in Human Behavior* 160 (2024) 108352. URL: <https://www.sciencedirect.com/science/article/pii/S0747563224002206>. doi:<https://doi.org/10.1016/j.chb.2024.108352>.
- [18] S. G. Hart, L. E. Staveland, Development of nasa-tlx (task load index): Results of empirical and theoretical research, in: P. A. Hancock, N. Meshkati (Eds.), *Human Mental Workload*, volume 52, North-Holland, 1988, pp. 139–183. URL: [https://doi.org/10.1016/S0166-4115\(08\)62386-9](https://doi.org/10.1016/S0166-4115(08)62386-9). doi:10.1016/S0166-4115(08)62386-9.
- [19] H. Beyer, K. Holtzblatt, *Contextual Design: Defining Customer-Centered Systems*, Morgan Kaufmann, San Francisco, CA, 1998.
- [20] F. D. Davis, Perceived usefulness, perceived ease of use, and user acceptance of information technology, *MIS quarterly* 13 (1989) 319–340.
- [21] V. Braun, V. Clarke, Using thematic analysis in psychology, *Qualitative Research in Psychology* 3 (2006) 77–101. URL: <https://doi.org/10.1191/1478088706qp063oa>. doi:10.1191/1478088706qp063oa.
- [22] E. F. Risko, S. J. Gilbert, Cognitive offloading, *Trends in Cognitive Sciences* 20 (2016) 676–688. URL: <https://doi.org/10.1016/j.tics.2016.07.002>. doi:10.1016/j.tics.2016.07.002.
- [23] A. Loukina, N. Madnani, K. Zechner, The many dimensions of algorithmic fairness in educational applications, in: *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 2019, pp. 1–10. URL: <https://doi.org/10.18653/v1/W19-4401>. doi:10.18653/v1/W19-4401.
- [24] R. Wang, R. Cheng, D. Ford, T. Zimmermann, Investigating and designing for trust in ai-powered code generation tools, in: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*, Association for Computing Machinery, 2024, pp. 1475–1493. URL: <https://doi.org/10.1145/3630106.3658984>. doi:10.1145/3630106.3658984.
- [25] Y. Zhang, Q. V. Liao, R. K. E. Bellamy, Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making, in: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT '20)*, Association for Computing Machinery, 2020, pp. 295–305. URL: <https://doi.org/10.1145/3351095.3372852>. doi:10.1145/3351095.3372852.
- [26] S. S. Y. Kim, Establishing appropriate trust in ai through transparency and explainability, in: *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '24)*, Association for Computing Machinery, 2024, pp. 1–6. URL: <https://doi.org/10.1145/3613905.3638184>. doi:10.1145/3613905.3638184.

A. Programming Exercise and Rubric

This appendix contains the materials used during the experiment. Figure 5 shows the exercise handout given to participants, and Table 3 shows the rubric breakdown.

Exer 03: Iterative Statements

INSTRUCTION: Create a Python program that will solve the problem below. Save the file as <Surname>_ex3.py (e.g. Trese_ex3.py). Submit your work on Google Classroom and turn it in. Write your name, section, and code description on the upper part of your document. Remember to document your work.

PROBLEM:

- A. Nested Loops and Infinite Loops** - Create a “menu” that infinitely loops until the user types 0. The menu has three choices, which are the two other parts of this exercise.

```
Welcome to this menu! If you think this works, you already have done half the battle!
(1) Do Part 1
(2) Do Part 2
(3) Do Bonus
(0) Exit application

Enter your choice: █
```

- B. Sequential Loops** – Using only a sequential loop (not nested), create the python code that prints the numbers from 1 to n^2 in a square matrix. See the example below when n is 5.

```
Enter your choice: 1
This is a for loop that prints from n to n^2, in an nxn pattern.
Enter a number: 5

    1    2    3    4    5
    6    7    8    9   10
   11   12   13   14   15
   16   17   18   19   20
   21   22   23   24   25
```

- C. Infinite Loops and Conditional Statements** - Create a “password verification” system that checks if the password is valid.

```
Welcome to this menu! If you think this works, you already have done half the battle!
(1) Do Part 1
(2) Do Part 2
(3) Do Bonus
(0) Exit application

Enter your choice: 2
Enter your password here!password
Please login your password here.dasd
Incorrect password!
Please try again.password
Incorrect password!
Please try again.password
Successful login!
```

- D. Bonus:** Print out a list of only even numbers between 1 to 100 without using a conditional statement. (worth 1 point)

Figure 5: Exercise handout.

Table 3
Rubric breakdown.

Functionality	Score
Main Menu	2
NxN Matrix	4
Password Authentication	3
Documentation	1
Bonus	1
TOTAL	11 pts max (10 base + 1 bonus)

B. ChatGPT Prompt Used in Grading Task

The following three prompts were used to guide ChatGPT in assessing the programming submissions. The first prompt established context, with the exercise handout shown in Appendix A attached:

As a computer science teacher in a university, I want you to grade the following exercises attached. I will send first the exercise handout given to the students which includes the instruction and expected deliverables of the Python program. Here are some notes that might help you consider the context:

- 1) The topic of this exercise is iteration.*
- 2) You may assume that they already knew Variables, Expressions, and Sequential Statements, and selection statements, and iterations only.*
- 3) The course where they did the exercise is Foundations of Computer Science (the first programming course the students will experience).*

I will send the submissions next and the rubrics.

The second prompt provided the rubric and granted permission to award partial credit. It also clarified that documentation should be evaluated without considering student names, which had been removed for anonymity:

Let me give you the rubric first:

Main Menu – (2pts)
NxN Matrix – (4pts)
Password Auth – (3pts)
Documentation – (1)
Bonus – (+1)

A total of 10 points overall (+1 if they successfully did the bonus part).

For every exercise file submission, I want you to show me the breakdown of the points in the rubric, as well as feedback in the given submission. All submissions are in Python code. You are also free to give partial points if you think they deserve to get one.

Kindly disregard the instruction to write their name and section. Only assess the code descriptions they have for the documentation.

The final prompt was reused for each submission, with the submission number substituted in:

Attached is submission X. Give me your assessment and feedback.

All prompts were issued to the o4-mini-high model via the ChatGPT web interface. Outputs were captured as static text and shown to participants in identical form across all sessions.

C. NASA-TLX Survey Format

Participants rated each of the following six NASA-TLX dimensions [18] on a scale from 0 (Very Low) to 7 (Very High):

1. Mental Demand: How mentally demanding was the grading task?
2. Physical Demand: How physically demanding was the task?
3. Temporal Demand: How hurried or rushed did you feel while performing the task?
4. Performance: How satisfied were you with your performance in the task?
5. Effort: How hard did you have to work to accomplish your level of performance?
6. Frustration: How insecure, discouraged, irritated, or annoyed did you feel during the task?

Two separate surveys were administered: one after the manual grading task and one after the ChatGPT-assisted grading task.

D. Contextual Inquiry and Post-Task Interview Questions

The contextual inquiries and post-task interviews followed a semi-structured format; additional questions were asked based on what each participant raised during the session. The prepared question set is reproduced below.

D.1. Feedback Quality

1. How would you describe the quality of the feedback provided by ChatGPT?
 - a) (Follow-up) Was it helpful in understanding the student's code or identifying mistakes?
2. Were there aspects of the ChatGPT feedback that you found particularly useful or clear?
 - a) (Follow-up) Can you give a specific example?
3. Were there aspects that you felt were lacking, misleading, or inaccurate?
 - a) (Follow-up) How did you handle these when reviewing the feedback?

D.2. Trust in ChatGPT

1. To what extent did you trust the grade and feedback generated by ChatGPT?
 - a) (Follow-up) Were there specific cases where you agreed or disagreed strongly?
2. Would you feel confident in using ChatGPT's feedback as-is, or would you prefer to always review and edit it?
 - a) (Follow-up) Why?
3. Do you think ChatGPT's feedback aligned with how you interpret the grading rubric?
 - a) (Follow-up) What would help improve that alignment?
4. From 0–100%, how would you rate your trust in using ChatGPT for grading? (Responses to this item produced the self-reported trust values discussed in Section 4.3.2.)
 - a) (Follow-up) Explain why.

D.3. Willingness to Adopt

1. Would you consider using ChatGPT or a similar tool in your own grading tasks in the future? Why or why not?
2. What would encourage you to adopt an AI-based grading assistant?
 - a) (Follow-up) What concerns or limitations would prevent you from using it regularly?
3. Do you think ChatGPT would be more useful for certain types of programming assignments (e.g., beginner-level vs. advanced)?
 - a) (Follow-up) Are there tasks where you feel AI should not be used?

D.4. Design Reflection

1. If you could redesign ChatGPT's grading behavior or interface, what features or safeguards would you add?