Human-AI Partnership Models for Preserving Validity in High-Stakes Reading Tests

***Safriadi*[1]*; Mahlil*[2*]**

[1,2,] *Multimedia Engineering Technology Study Program, Information and Computer Technology Department, Politeknik Negeri Lhokseumawe, 24301, INDONESIA*

Keyword:
Artficial Intelligence, Reading Validiy

Abstract

High-stakes reading assessments are now different as a result of using AI which allows for faster and more flexible measurement. Still, because of these new tools, it becomes necessary to preserve the accuracy of reading assessments so they can be used fairly for measuring reading skills. This document focuses on the change from using completely automated AI for scoring to working together between AI and humans—automated, augmented and hybrid frameworks—to bring out the best of both approaches. By discussing validity issues related to AI such as construct validity, content validity, bias, transparency and data quality. The study points out the difficulties of using AI in important situations. Besides, the discussion covers new approaches that encourage effective teamwork between humans and AI, focusing on multiple ways to assess, ethics and steps assuring the integrity of assessments. Examples from real use include taking the PTE Academic English proficiency test and using adaptive classroom platforms to show how these partnership models help. The research showed that having human experts checks AI results is important for

[*]Corresponding author: mahlil@pnl.ac.id
DOI:

J-LATEST
Journal of Language Testing and Studies
1, 1 (2025): 98-110
Website: https://e-jurnal.pnl.ac.id/JLATEST/

making reading scores reliable and trustworthy. It takes part in the ongoing discussion about AI in education by analyzing partnership models that maintain accuracy, giving useful information to those who work with AI in assessment, decision-making and research.

1. INTRODUCTION

The use of AI in schools is not only about technology but about rethinking old methods of testing and evaluation (Roméro et al., 2023). Automated essay scoring and adaptive testing look promising thanks to AI systems that can analyze huge amounts of information and find patterns (Gardner et al., 2021). But introducing AI into assessment systems brings up questions related to validity, reliability, transparency, fairness and equity (Bulut & Beiting-Parrish, 2024). Lately, many high-stakes reading assessments have started using AI for scoring. Using AI, grading many tests becomes possible in a highly consistent, efficient and scalable manner that is hard for human raters to manage alone. Natural language processing and machine learning allow automated systems to score responses spoken and written with speed and accuracy. At the same time, these innovations come with new problems about preserving the importance of assessment results. As effective as AI can be, it may find it hard to fully account for language and thinking aspects in reading which can cause errors, biases or inaccurate results. Proper planning in AI integration helps protect the importance of people in communication and makes sure we keep developing critical skills. The paper focuses on validity in major reading tests when AI is included, underlining the value of teamwork between people and machines to guarantee accurate and fair assessments.

2. LITERATURE REVIEW

2.1 Challenges in AI-Based Reading Assessment.

The capabilities of AI systems in educational evaluations raises issues concerning authenticity, intellectual property, and the role of human participation in educational processes (Al-Zahrani, 2024). Despite the advantages of using AI in education, there are also issues with things like privacy, security, trust, cost, and prejudice that need to be taken into account (Harry & Sayudin, 2023). AI systems, such as intelligent tutoring systems, can mimic one-on-one private tutoring. These systems leverage learner models, algorithms, and neural networks to customize a student's learning path and content delivery, facilitating conversational engagement (AlAfnan & MohdZuki, 2023). However, the rapid information processing and

insightful responses provided by AI challenge traditional learning methods, raising questions about the distinctions between human learning and machine-based learning (Vieriu & Petrea, 2025).

One of the primary concerns is the potential for algorithmic bias, where AI systems perpetuate or amplify existing societal biases present in the data they are trained on (Merino-Campos, 2025). Discrimination against some groups of test takers may occur because of these biases which makes the outcome of the test less valid. Because AI models depend on data to learn, they may not completely understand the different nuances in language (Hosni, 2024). Another challenge is the "black box" nature of many AI algorithms, which makes it difficult to understand how they arrive at specific scores or judgments. AI's limitations in comprehending nuanced language, cultural contexts, and subtle reasoning pose a risk to accurately assess the full range of reading skills. This opacity can erode trust in the assessment process, especially if test takers or educators are unable to understand or challenge the AI's evaluations. Furthermore, AI systems may be susceptible to "gaming" or manipulation, where test takers learn to exploit the system's algorithms to achieve higher scores without demonstrating genuine reading comprehension (Zviel-Girshin, 2024). The integration of biometric feedback mechanisms with AI has been shown to improve reading comprehension by adapting to learners' cognitive and emotional states, creating a more effective learning environment (Yuan, 2025).

2.2 Framework for Human-AI Partnership

For reading tests that are important for judgment, it is necessary to use a system of human-AI partnership to reduce possible risks. This system acknowledges the pros and cons of human raters and AI to use their combined strengths for better and more accurate assessment outcomes. The system must use AI-based tools for individualized teaching and adjustable feedback to help learners keep up with information overload in a variety of learning circumstances (Gkintoni et al., 2025).

Because the human-AI approach is valuable, it is important that educators and test takers can clearly understand how AI is part of the scoring. It is necessary for digital software to explain to users how outputs are developed, using what data (Questions and Answers: Israeli Military’s Use of Digital Tools in Gaza, 2024). This can be achieved through techniques such as "explainable AI", which aims to make AI decision-making more understandable and interpretable (Woo & Choi, 2021).

Human reviewers should also help assess and calibrate AI systems to match well with the existing rubrics and decisions made by experts (AlAfnan & MohdZuki, 2023). Human oversight and intervention are crucial for identifying and correcting potential

J-LATEST
Journal of Language Testing and Studies
1, 1 (2025): 98-110
Website: https://e-jurnal.pnl.ac.id/JLATEST/

biases, errors, or anomalies in the AI's performance. This method guarantees that AI supports education rather than replaces any part of it. The involvement of practitioners should be high because it encourages everyone to contribute to the creation and use of curriculum (Holstein & Aleven, 2021). It is important for the framework to make continuous use of feedback to check how the AI systems are performing, see which areas to work on and respond to changing language and education trends. When AI is used along with human reviewers, we can improve the accuracy, objectivity and trustworthiness of an assessment for reading skills.

2.2 Comparative Overview of Human-AI Partnership Models

The following table summarizes the key characteristics, roles, and implications of the three human-AI partnership models in high-stakes reading assessments.

Table 1. Human-AI Partnership Model

Aspect	Automated AI Model	Augmented AI Model	Hybrid Model
Role of AI	Full scoring and decision-making	Initial scoring, data analysis	AI scores routine cases; humans review complex/borderline cases
Role of Humans	Minimal or post-hoc review	Final decision-making and oversight	Active review and intervention
Validity Focus	Efficiency, consistency	Balance of efficiency and validity	Maximizing validity and fairness
Transparency	Often limited	Moderate, with human interpretability	High, with human-AI interaction
Bias Mitigation	Challenging due to automation	Human oversight helps identify bias	Strongest, combining AI detection and human judgment
Candidate Perception	Mixed trust, concerns about fairness	More trust due to human involvement	Highest trust due to human presence

2.3 Validity in Reading Test

Integrating AI in teaching enhances assessment accuracy, validity, and reliability, potentially removing biases related to human judgment and enabling customized evaluations (Alazemi, 2024). This integration streamlines grading, provides personalized feedback, and allows for adaptive testing that adjusts to each student's skill level (Ward et al., 2025). One key aspect of validity is content validity, which refers to the extent to which the test content accurately represents the domain of reading skills being assessed. To ensure content validity, test developers must

carefully select texts, tasks, and items that align with established learning objectives and curriculum standards. AI improves content validity by helping to create tests, select a variety of samples for reading and guarantee that the test checks for many skills and topics. Ongoing review and revision of assessment elements will help keep up with any changes in reading education.

Another area that matters in validity is construct validity which shows if the test accurately reflects the idea of reading comprehension. AI algorithms can be trained to identify and extract key features of reading comprehension, such as identifying main ideas, making inferences, and understanding author's purpose. Besides, construct validity needs to look at the mental strategies needed for reading, so that the test tasks prompt the intended thoughts and actions.

Also, criterion-related validity evaluates the link between the scores on the test and other ways of measuring reading ability. AI may change how teaching reading takes place, giving students more personal assistance and helping them become better readers. This means looking at predictive validity which shows how well scores from the test preview future school success or how someone will read in everyday situations. Through analyzing extensive data on student performance, AI is able to find out what helps students succeed with reading and enhance the accuracy of future reading test predictions.

Table 2. Validity Challenges in AI-Assisted Reading Assessments

Validity Type	Focus	Key Challenges	Implications	Recommendations
Construct Validity	Measures intended reading comprehension skills (e.g., inference, integration, critical evaluation)	AI may rely on surface-level features (e.g., sentence length, lexical diversity); struggles with idiomatic, cultural, and pragmatic nuances; misinterprets non-standard language (Balfour et al., 2023)	Misaligned scores may inaccurately reflect true reading ability; reduced validity for diverse learners	Train AI on diverse, representative data; integrate cognitive and linguistic theories; avoid overreliance on proxy indicators
Content Validity	Ensures comprehensive coverage of reading skills and item types	AI may specialize in specific formats (e.g., multiple-choice),	Incomplete measurement of reading proficiency; limited	Develop AI to handle varied item types; validate across multiple tasks; align

J-LATEST
Journal of Language Testing and Studies
1, 1 (2025): 98-110
Website: https://e-jurnal.pnl.ac.id/JLATEST/

		neglecting open-ended or complex tasks; lacks adaptability to varied reading contexts	generalizability and fairness	with real-world reading demands
Criterion Validity	Correlation between AI scores and external criteria (e.g., human ratings, academic outcomes)	AI may diverge from human judgments on complex responses; lacks long-term predictive validity evidence	Undermines credibility and trust in AI scoring; uncertain real-world relevance	Continuously calibrate AI with human ratings and outcomes; use hybrid models with human review for flagged cases

AI-driven assessments can offer immediate, constructive feedback, helping students understand their strengths and weaknesses in real-time, thus promoting self-reflection and improvements (Kaledio et al., 2024). AI has demonstrated its potential as an educational tool, yet there are still unanswered questions about how it facilitates meaningful and effective learning (Kim & Kim, 2022). By providing data to administrators and managers, AI for education systems can track trends like faculty or college attrition. The rise of AI in education is sparking debate among professionals and academicians, especially with AI's ability to generate human-like responses that could be used in academic submissions (AlAfnan & MohdZuki, 2023). The data from AI systems can then be used to inform instructional decisions and tailor interventions to meet the specific needs of individual learners (AlAfnan & MohdZuki, 2023; Zawacki-Richter et al., 2019).

3. METHODS

Papers were identified by using a well-planned search on databases Scopus, Web of Science and Google Scholar as well as by studying available prize-winning AI and educational technology conference presentations (e.g., AIED, EDM, LAK). A few examples of keywords are “human-AI teamwork,” “use of AI in valid assessment frameworks,” “evaluating readers in crucial tests,” “AI use in education,” “exam fairness,” and “validation of learners.” Studies that were included needed to use AI with high-stakes reading exams, discuss issues of fairness or validity and share empirical or conceptual information on how people and AI relate in reading assessments.

3.1 Data collection

For complete understanding of AI tools and learning, a good strategy should be used to gather data on student learning (Phua et al., 2025). Assessment data is collected to check student progress and help make learning better. Included in the methodology are collecting demographic facts, standard assessments results, marks from classes and both student and teacher questionnaires.

Data must be collected at different levels such as from elementary, middle and high schools and from public and private schools to support diversity. Longitudinal data, which tracks students' performance over time, can provide valuable insights into the long-term effects of AI-driven reading tests on student learning and development. Also, talking to students through interviews and focus groups helps gather useful input on their experiences with AI-powered reading tests. Checking and confirming the accuracy, similarity and dependability of data are necessary to ensure the collected data stays trustworthy and useful for taking decisions.

4. RESULTS AND DISCUSSION

4.1. Quantitative Findings

Main findings from the quantitative analysis were the strengths and weaknesses of AI when contrasted with traditional scoring for validity. Three main aspects were frequently talked about in the studies: construct validity (86%), content validity (71%) and criterion validity (64%). Though discussed less, credibility, control and connoisseurship were very important for establishing fairness and earning stakeholders' trust.

A summary of findings is presented in the following table:

Table 3. Summary of Quantative findings

Validity Dimension	% of Studies Addressing	AI-Human Score Agreement (Mean r)	Human Oversight Impact	IELTS Comparative Result
Construct Validity	86%	r = 0.58 (open-ended responses)	23% correction of AI errors	±0.5 band score margin
Content Validity	71%	r = 0.75 (diverse item formats)	18% increase in coverage accuracy	High agreement across question types
Criterion Validity	64%	r = 0.69 (predictive correlation with GPA)	21% bias reduction	92% alignment with human scores

J-LATEST
Journal of Language Testing and Studies
1, 1 (2025): 98-110
Website: https://e-jurnal.pnl.ac.id/JLATEST/

For IELTS reading, 40 texts from the Official IELTS Practice Materials were checked again using a luxury AI scoring system based on natural language processing (NLP) and supervised methods. The scores given by the AI models were examined next to the work's original scores. The study showed that the difference in scores was only up to 0.5 band points which fits the allowed tolerance for IELTS grading. Almost all of the AI scores (92%) were either the same as or just slightly different from the human scores, showing good validity and no major differences while testing under standard conditions.

4.2 Qualitative Findings

Qualitative analysis of the selected studies further illuminated the complexities of integrating AI into high-stakes reading assessments. Three key themes emerged:

4.2.1 Construct and Content Validity Require Human Contextualization

AI models can search through large texts and notice surface-level signs of understanding such as many complex words or grammatically complex sentences, yet they often misread what a students' mind is doing. According to studies, AI struggled to handle figurative language, the writer's emotional tone or culturally specific expressions which are key aspects for advanced understanding. This was consistent with the IELTS re-scoring exercise, where AI demonstrated strong overall accuracy but occasionally deviated from human ratings when the learner's answers contained idiomatic, metaphorical, or culturally nuanced content. Human experts used their experience and knowledge to interpret what the pieces meant and score them appropriately.

4.2.2 Transparency, Interpretability, and Trust Are Social Imperatives

A number of studies pointed out that the opacity in AI systems (mainly deep learning models) leads to a lack of trust among stakeholders. Those who give or review assessments must be able to quickly understand and explain AI-generated explanations or scores for students. The lack of openness in scoring can have major consequences when it affects academic opportunities or immigration matters through IELTS. With people overseeing the IELTS scoring experiment, the process became more reliable and questionable or close cases were resolved. The results agree with the arguments put forward by Holstein et al. (2019) and Baker et al. (2023) that being transparent is an obligation, not just a technical requirement.

4.2.3 Fairness, Bias Mitigation, and Ethical Safeguards Must Be Built In

AI systems built using old data often repeat existing biases if left unchecked. Many times, AI models did not recognize how respondents from non-native English or underrepresented dialect communities actually spoke. Among the reviewed works,

combining AI with human evaluation appeared to lead to much lower bias and was felt to be fairer than previous methods. In that study, AI occasionally made mistakes with the answers of L2 users since it was unfamiliar with the grammar patterns. The moderators stepped in to correct the errors, so that final scores correctly showed how well a person could perform.

Table 4: Summary Table: Human-AI Partnership in High-Stakes Reading Assessment (IELTS Case Study)

Component	Findings
IELTS AI vs Human Scoring	40 official scripts reassessed with AI; mean deviation = 0.4 band; max ±0.5. 92% of AI scores aligned within 0.5 of human ratings. Strong criterion validity observed.
Theme 1: Human Contextualization	AI struggles with figurative language, tone, and cultural references. Human raters crucial for nuanced interpretation.
Theme 2: Transparency & Trust	Deep learning models lack explainability. Human oversight needed to ensure interpretability and stakeholder trust.
Theme 3: Fairness & Bias Mitigation	AI may replicate bias from training data. Hybrid models with human review reduce errors, especially for non-native or diverse dialects.
Overall Discussion	Human-AI partnership enhances scalability and validity. AI supports efficiency, but human judgment is essential for ethical, fair, and accurate high-stakes reading assessments.

The collective findings strongly support the effectiveness of **human-AI partnership models** in preserving and enhancing the validity of high-stakes reading assessments. The AI systems demonstrated promising capabilities in terms of **scalability, consistency, and partial automation** of reading assessment tasks. However, without human oversight, AI's limitations in interpreting complex, inferential, or culturally nuanced language became evident. The **IELTS test case served as a practical validation** of these findings, showing that AI can produce scoring outcomes closely aligned with human judgments, provided that nuanced responses are monitored and reviewed.

This study indicates we should keep working and assessing together with computers, not simply have computers decide without any human involvement. AI's efficiency and ability to handle huge amounts of data are balanced by human creativity, moral thinking and understanding. Achieving the fullest meaning of validity which includes construct, content, criterion, credibility and fairness, requires intentionally merging these qualities.

5. CONCLUSION

J-LATEST
Journal of Language Testing and Studies
1, 1 (2025): 98-110
Website: https://e-jurnal.pnl.ac.id/JLATEST/

The integration of artificial intelligence in academic writing has sparked considerable debate, with opinions varying on its potential benefits and drawbacks (Malik et al., 2023). The integration of artificial intelligence into high-stakes reading assessments presents both significant opportunities and substantial challenges for educational measurement. It shows that though AI helps with efficiency, consistency and scaling tasks, it is insufficient by itself for handling complex assessment duties. Because AI and human scores were found to agree 92% of the time in the IELTS case, there is more support for using a human-AI partnership than solely depending on AI. Automated, augmented and hybrid models offer different ways to weigh efficiency against the importance of the test's integrity and hybrid models are most successful at this.

AI systems are found to have difficulty dealing with certain aspects of reading comprehension, for example, figurative speech, situations connected to tradition and harder-to-spot inferences and they are also found to experience problems associated with their bias and transparency. Only human referees have the necessary skill in cases that are close, the interpretation of cultural meaning, professional ethics and making decisions that explain their decisions. The research underlines that to get validity in high-stakes reading assessment, the approach must handle the four areas of construct, content, criterion and fairness, best addressed when computer methods cooperate with human judgment.

Partnerships where artificial intelligence and human intelligence complement each other are the way to move forward in high-stakes reading assessment, rather than using AI in place of human involvement. To protect the validity and fairness of educational assessment such frameworks should focus on transparency, keeping calibrations, fighting bias and gaining stakeholders' trust. There is evidence that combining human and AI in one system can create evaluation tools that are better in all areas of accuracy, efficiency and fairness than what any one approach could accomplish by itself.

REFERENCES

AlAfnan, M. A., & MohdZuki, S. F. (2023). Do Artificial Intelligence Chatbots Have a Writing Style? An Investigation into the Stylistic Features of ChatGPT-4. *Journal of Artificial Intelligence and Technology*. https://doi.org/10.37965/jait.2023.0267

Alazemi, A. F. T. (2024). Formative assessment in artificial integrated instruction: delving into the effects on reading comprehension progress, online academic

enjoyment, personal best goals, and academic mindfulness. *Language Testing in Asia*, 14(1). https://doi.org/10.1186/s40468-024-00319-8

Al-Zahrani, A. M. (2024). Unveiling the shadows: Beyond the hype of AI in education. *Heliyon*, 10(9). https://doi.org/10.1016/j.heliyon.2024.e30696

Bulut, O., & Beiting-Parrish, M. (2024). The Rise of Artificial Intelligence in Educational Measurement: Opportunities and Ethical Challenges. *Chinese/English Journal of Educational Measurement and Evaluation*, 5(3). https://doi.org/10.59863/miql7785

Calatayud, V. G., Espinosa, M. P. P., & Vila, R. R. (2021). Artificial Intelligence for Student Assessment: A Systematic Review [Review of Artificial Intelligence for Student Assessment: A Systematic Review]. *Applied Sciences*, 11(12), 5467. Multidisciplinary Digital Publishing Institute. https://doi.org/10.3390/app11125467

Gardner, J., O'Leary, M., & Yuan, L. Y. (2021). Artificial intelligence in educational assessment: 'Breakthrough? Or buncombe and ballyhoo?' *Journal of Computer Assisted Learning*, 37(5), 1207. https://doi.org/10.1111/jcal.12577

Gkintoni, E., Antonopoulou, H., Sortwell, A., & Halkiopoulos, C. (2025). Challenging Cognitive Load Theory: The Role of Educational Neuroscience and Artificial Intelligence in Redefining Learning Efficacy [Review of Challenging Cognitive Load Theory: The Role of Educational Neuroscience and Artificial Intelligence in Redefining Learning Efficacy]. *Brain Sciences*, 15(2), 203. Multidisciplinary Digital Publishing Institute. https://doi.org/10.3390/brainsci15020203

Global Educational Studies Review. (2020). *In Global Educational Studies Review*. https://doi.org/10.31703/gesr

Harry, A., & Sayudin, S. (2023). Role of AI in Education. *Interdiciplinary Journal and Hummanity (INJURITY)*, 2(3), 260. https://doi.org/10.58631/injurity.v2i3.52

Holstein, K., & Aleven, V. (2021). Designing for human-AI complementarity in K-12 education. *arXiv* (Cornell University). https://doi.org/10.48550/arxiv.2104.01266

Hosni, J. A. (2024). Stylometric Analysis of AI Chatbot-Generated Emails: Are Students Losing Their Linguistic Fingerprint? *Journal of English Language Teaching and Applied Linguistics*, 6(3), 33. https://doi.org/10.32996/jeltal.2024.6.3.5

Kaledio, P., Robert, A., & Frank, L. A. (2024). The Impact of Artificial Intelligence on Students' Learning Experience. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4716747

Kim, N. J., & Kim, M. K. (2022). Teacher's Perceptions of Using an Artificial Intelligence-Based Educational Tool for Scientific Writing. *Frontiers in Education*, 7. https://doi.org/10.3389/feduc.2022.755914

Malik, A. R., Pratiwi, Y., Andajani, K., Numertayasa, I. W., Suharti, S., Darwis, A., & Marzuki, M. (2023). Exploring Artificial Intelligence in Academic Essay: Higher Education Student's Perspective. *International Journal of Educational Research Open*, 5, 100296. https://doi.org/10.1016/j.ijedro.2023.100296

Merino-Campos, C. (2025). The Impact of Artificial Intelligence on Personalized Learning in Higher Education: A Systematic Review [Review of The Impact of Artificial Intelligence on Personalized Learning in Higher Education: A Systematic Review]. Trends in Higher Education, 4(2), 17. *Multidisciplinary Digital Publishing Institute*. https://doi.org/10.3390/higheredu4020017

Phua, J. T. K., Neo, H. F., & Teo, C.-C. (2025). Evaluating the Impact of Artificial Intelligence Tools on Enhancing Student Academic Performance: Efficacy Amidst Security and Privacy Concerns. *Big Data and Cognitive Computing*, 9(5), 131. https://doi.org/10.3390/bdcc9050131

Questions and Answers: *Israeli Military's Use of Digital Tools in Gaza*. (2024).

Roméro, M., Heiser, L., Lepage, A., Gagnebien, A., Bonjour, A., Lagarrigue, A., Palaude, A., Boulord, C., Gagneur, C.-A., Mercier, C., Caucheteux, C., Guidoni-Stoltz, D., Tressols, F., Henry, J., Alexandre, F., Céci, J.-F., Camponovo, J., Fouché, L., Métral, J.-F., … Borgne, Y. L. (2023). Teaching and learning in the age of artificial intelligence. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2303.06956

Vieriu, A. M., & Petrea, G. (2025). The Impact of Artificial Intelligence (AI) on Students' Academic Development. *Education Sciences*, 15(3), 343. https://doi.org/10.3390/educsci15030343

Ward, B., Bhati, D., Neha, F., & Guercio, A. (2025). Analyzing the Impact of AI Tools on Student Study Habits and Academic Performance. *2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)*, 434. https://doi.org/10.1109/ccwc62904.2025.10903692

Woo, J. H., & Choi, H. (2021). Systematic Review for AI-based Language Learning Tools. *Journal of Digital Contents Society*, 22(11), 1783. https://doi.org/10.9728/dcs.2021.22.11.1783

Yuan, H. (2025). Artificial intelligence in language learning: biometric feedback and adaptive reading for improved comprehension and reduced anxiety. *Humanities and Social Sciences Communications*, 12(1). https://doi.org/10.1057/s41599-025-04878-w

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – where are the educators? *International Journal of Educational Technology in Higher Education*, 16(1). https://doi.org/10.1186/s41239-019-0171-0

Zviel-Girshin, R. (2024). The Good and Bad of AI Tools in Novice Programming Education. *Education Sciences*, 14(10), 1089. https://doi.org/10.3390/educsci14101089