Breakpoints in iterative development and interdisciplinary collaboration of AI-driven automated assessment

Xiaoshan Huang, Department of Teacher Education, University of Turku, xihuan@utu.fi

Li-Hsin Chang, Department of Computing, University of Turku, lhchan@utu.fi

Koen Veermans, Department of Teacher Education, University of Turku, koevee@utu.fi

Filip Ginter, Department of Computing, University of Turku, figint@utu.fi

ABSTRACT

The rise of AI in education has led to significant advancements, promoting automated grading to reduce educator workload and to enhance pedagogy. However, its integration raises complex pedagogical, ethical, and technical questions. This systematic review examines the intersection of automated grading tool development and educational assessment through the lens of the activity theory. Our analysis, informed by literature since 2010, reveals a critical need for comprehensive evaluation frameworks addressing the iterative nature of technology development and interdisciplinary collaboration. Key breakpoints in existing studies include oversight of the reliability and validity of assessments, ethical considerations, coherent evaluation rules, interdisciplinary collaboration, and agentive and constructive roles of users. Addressing these issues requires a holistic approach that bridges technical and educational perspectives, fostering trust and supporting meaningful learning outcomes. Enhanced collaboration and ongoing professional development are crucial for creating AI-driven assessments.

Keywords: *Automated grading, Artificial intelligence, Assessment, Instructional design, Pedagogy*

INTRODUCTION

The rise of AI has gained global attention in the educational field, especially in assessments, where its applications span assessment for learning, evaluation, and predictive analytics [1]. his integration has driven the development of automated assessment systems designed to enhance and streamline assessment processes, aiming to improve pedagogical value and to reduce educators' workloads [2]. Automated assessments encompass varying levels of grading or scoring automation and provide informative evaluations and pedagogical functionalities such as generating personalized feedback, developing adaptive testing mechanisms, and conducting predictive analytics to assess and enhance student performance.

Given the foundational role of automated grading tools in these systems, this article conducts a systematic literature review to explore the multifaceted landscape of automated grading systems. The review examines the assessment activities and purposes these automated grading tools support, the evaluation indicators used to assess their effectiveness and fairness, and the interdisciplinary collaboration required for their development and deployment, by bridging educational and computer science perspectives.

Automated grading has evolved from simple, predefined scoring of multiple-choice questions to more sophisticated generative scoring methods like automated essay scoring (AES), automated writing evaluation (AWE), and automated short answer grading (ASAG). These advanced systems leverage AI capabilities, such as machine learning algorithms, to analyze large-scale, continuous, and multifaceted assessment data [3], [4]. Despite advancements, fundamental challenges recur in the reductionism inherent in automated grading, which may overlook higher-order thinking and text meaning, with an overemphasis on superficial features detracting from content quality [5], [6]. Another fundamental issue is to consider the validity and accuracy of automated grading merely based on human-machine agreement, which focuses on the final grading results but neglects the fact that human-machine agreement may correlate but does not necessarily reflect genuine alignment [4].

This highlights the distinction between technological innovation and the underlying theoretical assessment framework [7]. Such distinction raises different issues, including an emphasis on superficial features detracting from content quality, students gaming the system, and low accountability in high-stakes assessment contexts [8]. Efforts have been made to improve the performance of algorithmic scoring systems by optimizing technical factors such as data quality and quantity, feature engineering, algorithm selection, parameter tuning, and evaluation metrics [9], [10]. Nevertheless, challenges linger

in translating abstract and complex constructs into clear and reliable annotation guidelines for rubric items that can be used by machines [11].

Grading validity of automated grading tools should be designed as part of a construct-driven assessment system to ensure grades accurately reflect the construct being measured [12]. Scoring algorithms should align response features with targeted cognitive and learning constructs in specific contexts [5]. The evidence aligns the feature extraction, weighting, and evaluation with the targeted constructs and criteria [13]. Training data should help the model recognize purposeful diagnostic information, ensuring the data is understandable regarding the construct across proficiency levels. AI scores can be used alone or combined with human-scored responses or machine-scored items [13]. Effective validation of AI scoring should focus on the construct validity of algorithmic scoring, which is grounded in a robust integration of logic and data and provide a comprehensive analysis of how accurately the scores reflect the intended construct and their resistance to irrelevant variance [12].

The challenges of accuracy, reliability, and validity are not merely technical problems, but as they are deeply intertwined with complexities of human cognition, language, and ethical considerations. Early-stage collaboration between developers and educators to analyze in analyzing targeted knowledge and skills, tasks, and evidence is essential for creating valid, transparent, and interpretable variables, as well as models in algorithmic grading that align with specific pedagogical assessment purposes. Despite the necessity of collaboration, reviews of recent AI developments in education show a notable gap in the involvement of educational departments [6].

The first issue is the alignment of automated grading tools with assessment purposes. One purpose is the assessment of learning (AoL), which measures what students have learned at the end of a unit or course [14]. Another purpose is the assessment for learning (AfL), which involves seeking and interpreting evidence to decide where learners are in their learning process, where they need to go, and how best to get there with continuous feedback [15]. The purpose of assessment as learning (AaL), which encourages learners to take responsibility for their learning and play agentic and constructive roles in understanding the assessment, is often overlooked [14]. In assessment settings, learners and teachers should critically evaluate automated grading tools to have a voice in the assessment process and to enhance their meta-cognitive skills and self-regulation. The engagement of students and teachers are essential for the feedback loop for developing automated grading tools that align with educational purposes [16].

The gap in interdisciplinary collaboration has led to challenges concerning the educational, practical, and ethical values of AI-based assessments, particularly regarding bias, fairness, transparency, accountability, and the need for equitable assessment practices [17], [18]. It is essential to incorporate ethical principles at the beginning of development and application, especially in AI involvement. Principles such as justice and fairness, transparency, non-maleficence, responsibility, and privacy must be considered [19].

To advance the development and integration of automated grading tools in education, several key issues must be addressed: 1) synergies and conflicts between automated grading tools and educational assessment theories (accuracy, reliability, and validity), 2) the interdisciplinary and ethical development of AI-driven assessments, 3) the active roles of stakeholders in developing and using these tools. Therefore, conducting a comprehensive literature review on automated grading tools is essential to gain an in-depth understanding of the current assessment purposes and established rules of developing AI-driven automated grading tools, as well as current interdisciplinary collaborations and the obstacles hindering the development of ethical automated grading tools with educational value.

The systematic review requires a robust theoretical framework. Activity Theory (AT) is particularly well-suited for this purpose due to its holistic and socio-cultural approach to understanding human practices as systemic and socially situated phenomena [20], [21].

The basic unit of analysis in activity theory is activity, which incorporates human actions within meaningful contexts (see Figure 1). Activities involve sequences of individual and cooperative actions aimed at transforming an object, mediated by tools, into desired outcomes [21], [22]. The foundation of any activity lies within its community, governed by rules and the division of labor, which are products of historical development and subject to ongoing refinement. Engaging in an activity means executing actions with specific, immediate goals. The planning of these actions often relies on an action model; the more refined this model is, the more likely is the success of the ensuing action [22].

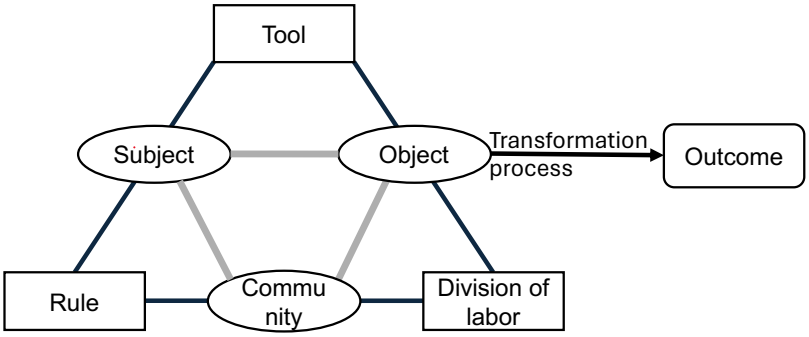

Figure *1* Basic structure of an activity, there are extra spaces and the number is in italics.

While activity theory identifies contradictions and tensions within an activity system effectively, such as conflicts between pedagogical objectives and the constraints of automated systems, it also recognizes the dynamic evolution of activities (i.e., the development of automated assessment). This adaptability aligns with the rapidly evolving field of AI and automated assessment, where interactions and community regulations change continuously. By employing the activity theory, we can systematically review the development and implementation of automated grading tools in educational contexts. The motive in existing literature reveals the driving forces and assessment activities of developing automated grading tools. The examination of the rules in different activity systems from educational and computer science fields allows to highlight the alignment or misalignment with technical focuses and educational assessment values. The analysis of the subjects and divisions of labor showcases interdisciplinary collaboration and stakeholder roles. Analyzing the automated grading tools used reveals their mediating function in grading, learning, and teaching.

Research questions

How are automated grading tools developed and integrated into educational assessment activities from an interdisciplinary perspective?

Sub-questions guiding the data analysis:

1. What are the motives underlying existing research on automated grading tools?
2. What are the assessment purposes of automated grading tools under different motives?
3. What are the rules, i.e., evaluation indicators, prioritized in computer and educational fields in these activities?
4. What are the action models in the activities driven by these motives?
5. Who are the subjects and what divisions of labor are involved in these activities?

METHOD

To gather relevant studies on the use of automated grading tools in educational environments, we followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [23]. Figure 2 illustrates the process of literature identification, screening, and selection for this review.

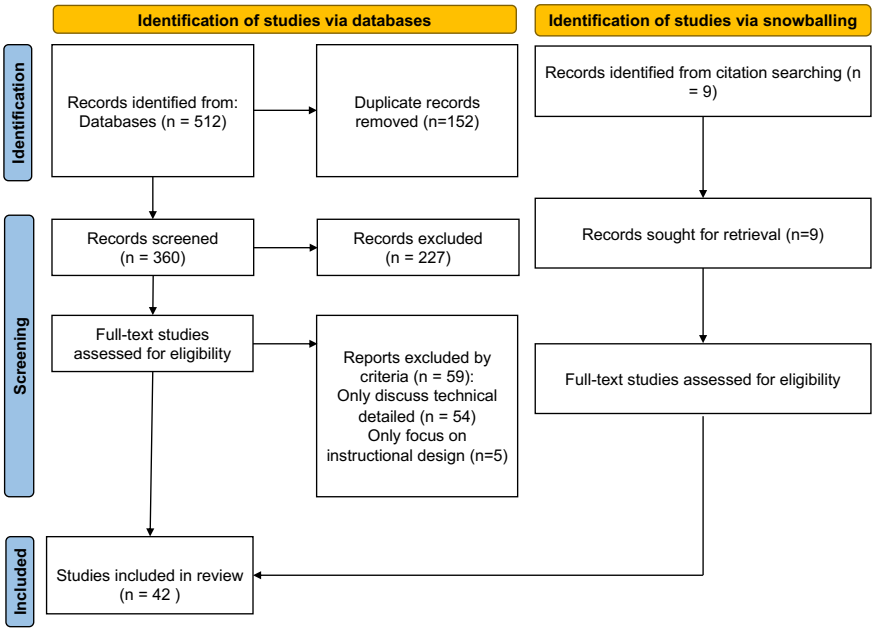

Figure 2 Selection procedure flow chart

Literature search

The search for relevant literature was conducted using two prominent electronic databases in the field of educational science: EBSCO and ProQuest. Search phrases such as “automat* grading,” “computer-aided grading,” “AI in grading,” “algorithmic grading,” “machine learning in grading,” and “technology-assisted grading” were employed. These phrases were combined using “OR” to broaden the search scope. Our systematic review focused on articles published since 2010, due to significant advancements in natural language processing (NLP), machine learning (ML), and artificial intelligence (AI) technologies in the early 2010s. Additionally, the review was limited to studies published in peer-reviewed journals and available in English. The search, executed in October 2023, identified 521 studies.

Literature screening

To examine the complex interplay between the development of automated grading tools and educational assessment activities, the included studies must provide sufficient information on the impact of these tools on learning, teaching, grading, and their interaction with human grading. The literature screening ensures the studies meeting the following criteria 1) The research should provide information on the consequences of using automated grading tools, including their impact on students’ learning outcomes, instructors’ teaching and grading methods, or perceptions of automated grading tools, 2) The research should provide information on the interaction between automated grading tools and humans during learning, teaching, and grading processes, 3) The research should provide empirical evidence related to the content stated in criteria 1 and 2, meaning the research must involve specific automated grading tools.

The articles must meet one of the first two criteria and invariably fulfill the third criterion in order to be included. Articles focusing solely on the technical details of automated grading tools or instructional design without relevant analysis related to criteria 1 or 2 are excluded. We conducted the screening process using a double-blind method based on the titles and abstracts of the articles. For conflicting decisions and “maybe” articles, we reviewed the full text for comprehensive evaluation. Ultimately, 42 papers were included for analysis.

Data analysis

Data analysis was performed using both deductive and inductive coding (see Figure 3). We coded each included study as an individual activity within the domain of advancing automated grading in educational activity system, totaling 42 activities. For deductive coding, guided by activity theory, we segmented the text in each included study to describe 1) motives: text describing the aims and objectives of the activity (included study), 2) subjects: text describing the individual or group interacting with the automated assessment tools and involved in the assessment activities, 3) action models:text describing the chain of actions taken to address the research questions or achieve the activity goals, 4) rules: text describing the evaluation indicators used to mediate the subjects' actions and evaluate the automated assessment

tools, 5) division of labor: text describing the distribution of tasks, responsibilities, roles, and power among the subjects involved in the activity,6) tools: text describing the automated assessment system, and 7) outcomes: text describing the results of the activity. To capture the complexities, texts describing key information were segmented in NVivo.

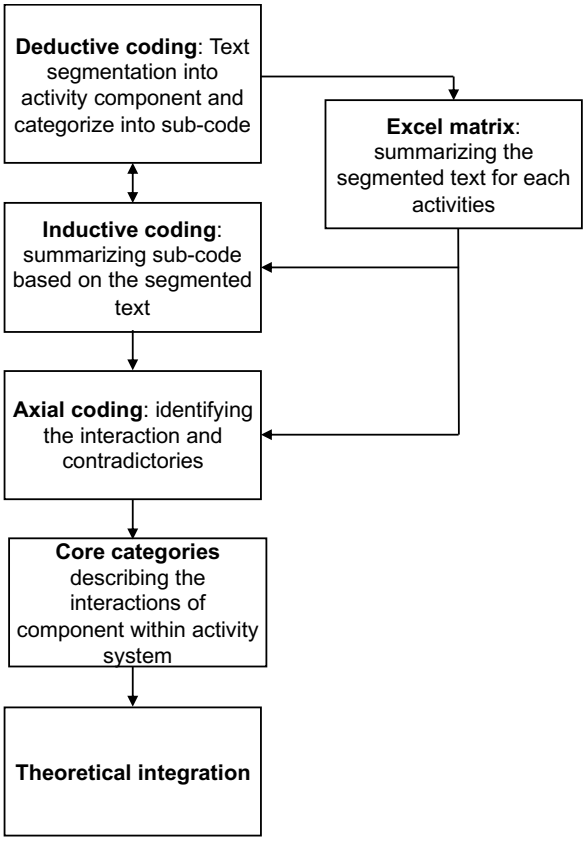

Figure 3 Thematic data analysis process

For inductive coding, themes emerging from segmented texts were identified as sub-codes of the seven main codes. Periodic meetings were held to discuss and interpret these patterns. Emerged patterns and sub-themes were iteratively used for further deductive coding if aligned with existing theories. For instance, analyzing the assessment purposes of automate grading tools revealed themes like formative assessment for learning and summative assessment of learning, with one case showing critical pedagogy in using assessment as learning. We identified three main categories: assessment of, for, and as learning.

RESULTS

What are the motives of activities, i.e. included studies, on automated grading tools?

The findings indicate two main categories of motives in the activities (i.e., included studies). One activity could engage in more than one sub-motive. Motive 1 focuses on developing automated assessment tools by increasing their pedagogical and evaluative value and enhancing their technical capabilities (n=20). Sub-motive 1.1 aims to increase the pedagogical effectiveness of assessments by providing feedback to improve students' learning outcomes (n=7) [24], [25], [26], [27], [28], [29], [30]. Sub-motive 1.2 seeks to enhance the evaluative function by assessing complex tasks that elicit higher-order thinking and problem-solving (n=7) [30], [31], [32], [33], [34], [35], [36]. Sub-motive 1.3 advances technical possibilities by increasing data efficiency, improving grading reliability, and addressing ethical issues such as fairness, human-centric approaches, adaptability, and contextualization (n=6) [29], [37], [38], [39], [40], [41]. Initiatives under Motive 1 often emphasize alleviating the resource-intensive nature of traditional assessments, as well as addressing biases and inconsistencies in human grading.

Motive 2 focuses on evaluating the performance and impact of automated assessments (n=22). Sub-motive 2.1 evaluates validity and reliability (n=2) [42], [43]. Sub-motive 2.2 assesses the impact on learning outcomes (n=3) [44], [45], [46]. Sub-motive 2.3 examines users' perceptions, acceptance, and trust (n=9) [7], [8], [47], [48], [49], [50], [51], [52], [53]. Sub-motive 2.4 evaluates technical performance

(n=10) [7], [10], [54], [55], [56], [57], [58], [59], [60], [61]. Sub-motive 2.5 investigates user interaction (n=5) [7], [47], [51], [55], [62] (see Figure 4).

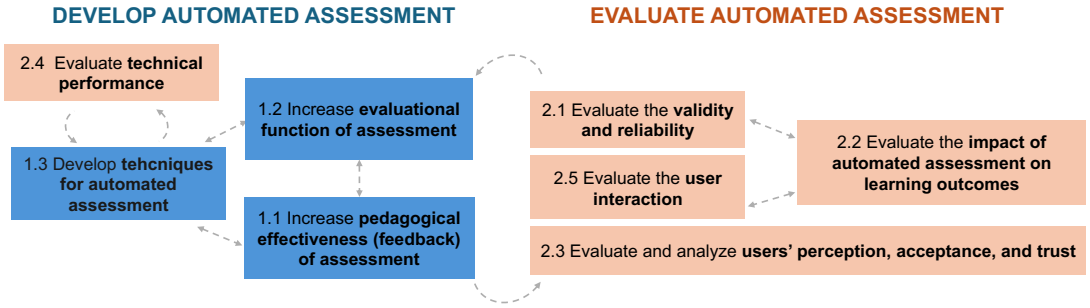

Figure 4 Overview of motives of included studies

Despite the fundamental importance of reliability and construct validity in automated grading, this area has received very negligible attention. This highlights a lack of understanding of the fundamental issues regarding the construct validity of automated grading tools and their alignment with educational values. There are limited number of studies on the interaction between users and automated assessment tools, suggesting a potential gap in understanding how learners and teachers play agentic and constructive roles in the use of these tools. Additionally, efforts to address the ethical issues of automated grading tools are relatively minimal.

The motives of the activities reveal interactions among individual activity systems that collectively form a larger activity system (see Figure 4). Theoretically, outcomes from Motive 1.3 lay the technical foundation for activities in Motives 1.1 and 1.2. Activities under Motive 1.2, aimed at enhancing evaluative functions for complex tasks, can also develop through Motive 1.1. Outcomes from Motive 1 activities should precede Motive 2 activities, providing feedback for future cycles of Motive 1 activities. This broader system demonstrates the iterative nature of developing these tools, where smaller activity systems interact to advance the overall field of automated assessment in education. However, the imbalance in the number of activities with different motives indicates a disproportionate focus within the broader activity system, potentially disrupting the iterative and interdisciplinary development loop.

What are the assessment purposes of automated grading tools under different motives?

Activities focused on summative assessment of learning are considerably more than formative assessment for learning (see Figure 5). The activities emphasizing the pedagogical value of assessment (Motives 1.1, 2.2, 2.5) (n=20) [7], [24], [25], [26], [27], [28], [29], [30], [32], [37], [38], [45], [46], [47], [48], [50], [54], [59], [60], place a balanced emphasis on both summative and formative assessment. In contrast, activities focusing on the evaluative function (Motive 1.2), technical performance (Motives 1.3 and 2.4), validity and reliability (Motive 2.1), and user perception (Motive 2.3) place a heavier emphasis on summative assessment (n=22) [8], [10], [31], [33], [34], [35], [36], [39], [40], [41], [42], [43], [49], [51], [52], [53], [54], [56], [57], [58], [61], [62].

Theoretically, under the broader activity system, activities under Motives 1.3 and 2.4 provide the technical foundation. However, only a small number of activities have provided technical advancements

for formative assessments (n=7) [7], [29], [37], [38], [55], [59], [60], which points out a possible gap in the development of automated grading tools for formative assessment. This finding suggests that

automated grading tools primarily focus on summative assessment , whereas technical advancement focuses on formative assesment for learning.

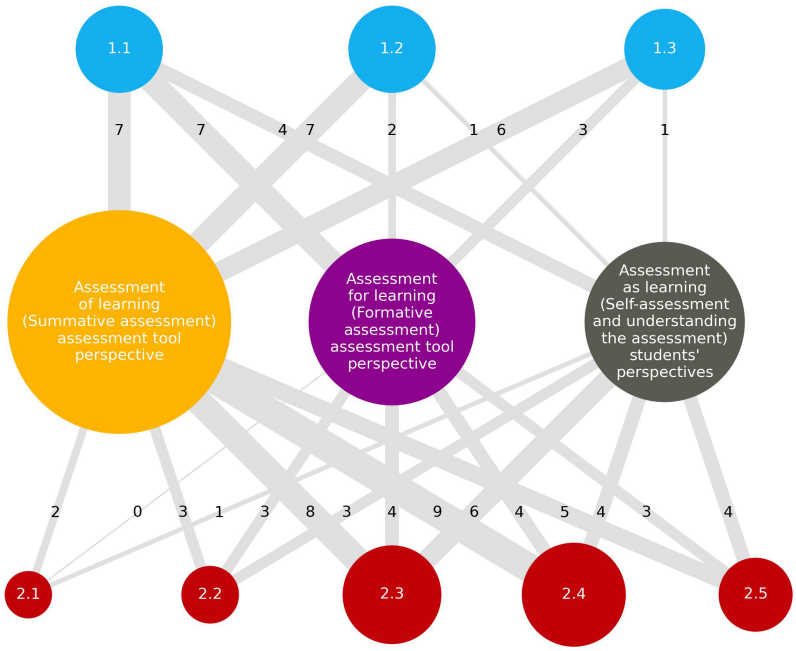

Figure 5 Overview of the activities' motives and purposes of automated assessments

Motive 1.1 = Develop automated assessment to increase pedagogical effectiveness of assessment (feedback)
Motive 1.2 = Develop automated assessment to increase evaluative function of assessment for complex tasks
Motive 1.3 = Develop automated assessment with high technical performance
Motive 2.1 = Evaluate the validity and reliability of automated assessment
Motive 2.2 = Evaluate the impact of automated assessment on learning outcomes
Motive 2.3 = Analysis and evaluate the perception, acceptance, and trust in automated assessment
Motive 2.4 = Evaluate the technical performance of automated assessment
Motive 2.5 = Evaluate the user interaction with automated assessment tool

We encoded assessment as learning as the involvement of students in understanding automated assessment tools, having the students reflect on their use and consider the tools's impact on learning processes and outcomes (n=17) [7], [8], [24], [25], [29], [30], [43], [44], [45], [46], [47], [50], [51], [53], [54], [60], [61](see Figure 5). Activities with Motives 1.1, 2.2, 2.3, and 2.5, which focus on learners' perspectives, active learning processes, and impacts on learning outcomes, involve more assessment as learning. In contrast, activities with Motives 1.2, 1.3, and 2.4, which focus more on assessment results and technical performance, involve less assessment as learning. This indicates that the engagement of learners as active users with clear learning objectives more naturally involves them in assessment-as-learning (there are not always hyphens between these terms, maybe add them to all?) activities compared to activities where learners play a passive role.

However, activities with Motives 1.3 and 2.4 show that some assessment as learning activities can still occur even when the focus is on technical performance (n=5) [7], [29], [55], [60], [61]. This indicates that it is possible to engage students in evaluating automated grading tools and reflecting their impact on learning, even within a technically focused context. More activities that involve learners in understanding the technical performance of automated grading tools and using summative grading tools constructively and proactively are needed.

What are the rules, i.e., evaluation indicators, prioritized in computer and educational field in these activities?

The evaluation indicators used in the included studies reveal fundamental differences in how automated grading tools are evaluated from technical and educational perspectives. To understand these priorities better, we categorized activities into educational focuses (motives 1.1, 1.2, 2.1, 2.2, 2.3, 2.5) and technical focuses (motives 1.3, 2.4). Qualitative data analysis indicates three groups of evaluation

indicators used to assess the performance, effectiveness, and impact of automated grading tools (see Figure 6): 1) technical perspective: technical performance, 2) Educational perspectives: users' evaluation and perception, and 3) impact on teaching, grading, and learning.

The results suggest that technical and educational perspectives might use the same terms with different meanings and expectations. Fundamentally, technical perspectives focus on validating tool performance, accuracy, and scalability, whereas educational perspectives emphasize impact on learning outcomes, alignment of tools with educational objectives, and stakeholder perceptions. The gap between technical and educational perspectives also appears when the impact on teaching, grading, and learning is measured. Technical perspectives focus on objective measurements such as grading time and error reduction. In contrast, educational perspectives focus more on teachers and learners actively reflecting on the impact, covering broader aspects of the influence of automated grading tools. While objective data shows clear benefits of understanding the performance of automated

grading tools, subjective experiences highlight inaccurate results, practical challenges, and the alignments with educational meaning.

Technical perspectives: **Technical performance**	**Educational perspectives:** **Users' (i.e. instructors and learners) evaluation and and percpetion**
Reliability, Validity, and Accuracy Scoring precision, error rate, inter-rater reliability (human grade as gold standard), consistency, 10-fold-cross validation.	**Perception on Reliability and Validity** Human-machine agreement, error rate, grading consistency of scores across multiple assessments, alignment with learning objectives, accuracy in capturing student response nuances
Explainability System's ability to provide clear, understandable, and interpretable reasons behind decisions or scores.	**Perception on Clarity (explainablity)** System's ability to provide clear, understandable, and interpretable feedback and reasoning behind decisions or scores.
Feasibility and Practicality Financial/resource investment, data quality, computational power, technical expertise.	**Perception on Usability and Practicality** Ease of use, user interface design, accessibility, intuitiveness, functionality, integration with existing systems, reliability, tangible benefits.
Fairness and Biases Fairness and biases from algorithms and datasets: ABROCA values, token recognition.	**Ethical Considerations** Transparency, fairness, human oversight, traceability, explainability, algorithm biases.
Other Technical Indicators	**Other Educational Indicator**
Adaptability Ease of updating/modifying system, handling different questions/tasks, integrating human collaboration.	**Attitude and Acceptance** Initial reactions, ongoing impressions, satisfaction scores, trust, engagement, preference vs. human feedback, resistance, concerns.
Speed and Scalability Handling large datasets/users, processing/grading time.	
Impact on teaching, grading, and learning	
Passively Measured Subjects	**Actively Reflective Subjects**
Impact on Teaching and Grading Time/effort on grading, grading performance improvements, objective/uniform assessments.	**Reflection of Impact on Teaching and Grading** Time management, job performance, feedback quality, economic benefits from tool implementation.
Impact on Learning Outcomes Grade improvements, error reduction, writing quality enhancement, learning approach changes.	**Reflection of Impact on Learning** Outcomes, engagement, study habits/strategies, effectiveness, exam preparation, motivation, anxiety/stress, impact on student admission.

Figure 6 Overview of evaluation indicators used in activities with motives of educational or technical focuses

Qualitative analysis of the evaluation indicators indicates different expectations of automated grading tools from different fields. Quantitative analysis indicates the overall focuses across two field (see Figure 7). The analysis reveals that technical indicators such as accuracy, reliability, and validity are the primary measures used to evaluate automated grading tools [7], [10], [26], [27], [29], [30], [33], [35], [36], [39], [40], [41], [42], [46], [48], [58], [59], [60], [61], [62]. However, validity assessments mainly rely on human-machine agreement, error rates, and 10-fold cross-validation, with little emphasis on the construct validity of the tools. These evaluations use statistical methods to identify patterns and conditions under which automated grading performs well or poorly. These evaluation only offer probabilistic accuracy that may decrease with diverse response structures or different contexts. Issues

like false negatives, false positives, and difficulties in rating certain response types persist despite high accuracy in general.

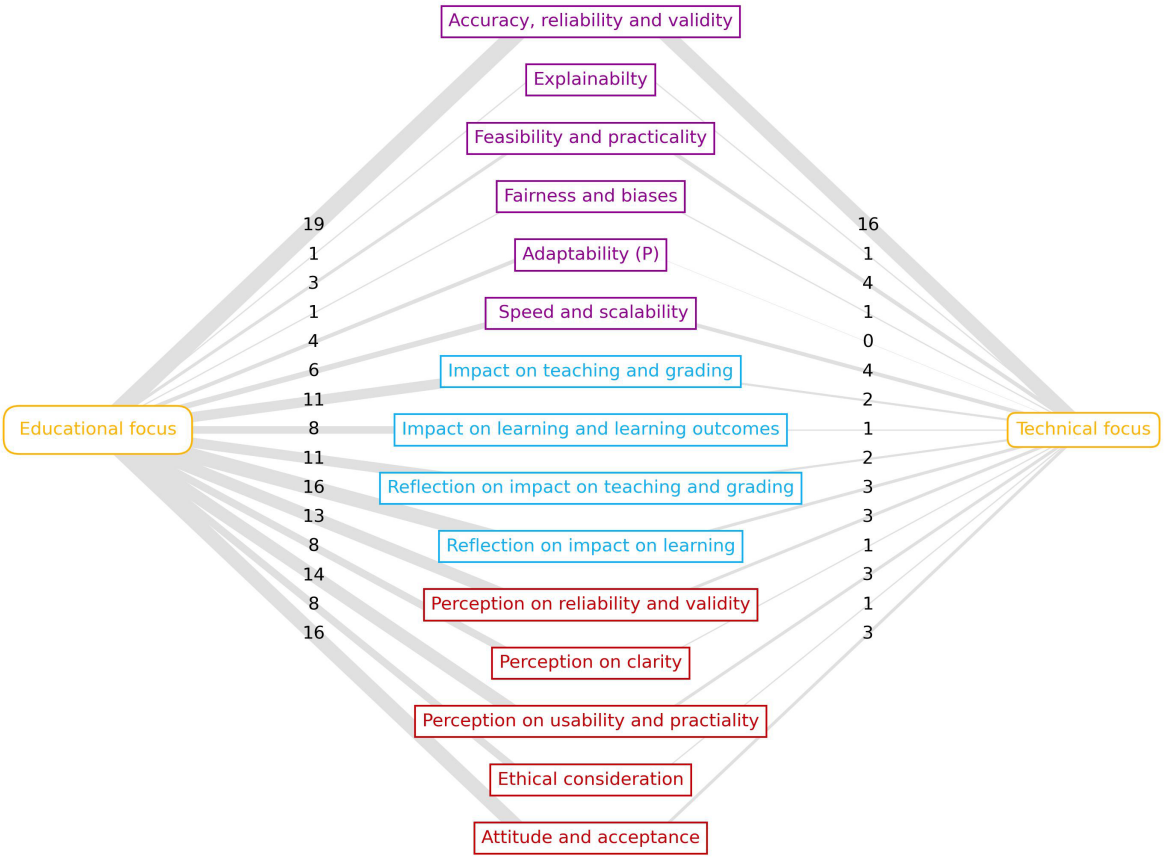

Figure 7 *Overview of evaluation indicators used in activities with motives of educational or technical focuses*

Furthermore, evaluation indicators like “explainability” [7], [8], [24], [28], [35], [47], [49], [51], “fairness and biases” [7], [25], [32], [34], [49], [52], [56] are underutilized in assessing automated grading tools. This under-utilization suggests a need for technical evaluations to focus on ethical AI principles. Similarly, indicators such as “ethical consideration” and “perception of clarity” are less frequently used, indicating that while these aspects are acknowledged, they are not primary focuses in both field. The lack of emphasis on ethics in both fields highlights the necessity to integrate these ethical elements more thoroughly into the evaluations of automated grading tools.

The results also indicate that educational activities encompass a broader range of user perceptions and experiences [7], [24], [25], [29], [34], [35], [46], [47], [48], [50], [53], [55], [62]. Additionally, objective measurements of the impact on teaching, grading, learning, and learning outcomes are underemphasize in technical focus activities when evaluating automated grading tools. This lack of focus may result in developing tools that lack real-world applicability and educational value.

What are the action models in the activities driven by these motives?

Activities with different motives demonstrated action models, a series of actions driven by motives, that align with the phases of the ADDIE model: Analysis, Design, Development, Implementation, and Evaluation (see Figure 8).

Our findings reveal discrepancies between activities with educational advancement focuses and those with technical advancement focuses. The activities aiming to develop automated assessment tools (motives 1.1, 1.2, 1.3) generally cover the full ADDIE cycle. However, in motive 1.3, activities focused on technical advancement have fewer instances in the implementation phase (n=3) (n=3) [29], [38], [40], in real-world educational contexts. This gap suggests that the evaluations of technical performance in these activities primarily rely on “process data” rather than real-world contexts [37], [39], [41].

Activities with motive 2 also exhibit discrepancies between technical and educational focuses. Activities emphasizing educational and stakeholder perspectives (motives 2.1, 2.2, 2.3, 2.5) are minimally involved in the design [55] and development[42], [46], [50], [54], [62] phases, instead focusing on implementing and evaluating automated assessment tools in real-world contexts. These activities often adopt existing tools but may lack refinement for specific educational purposes. In contrast, activities focusing on technical performance (motive 2.4) engage in both the design [57], [59] and development [10], [54], [55], [56], [57], [58], [59] phases. These activities prioritize refining algorithms and retraining models with better data or features to contextualize the usage of automated grading tools [7], [56], [60], [61].

These discrepancies highlight differences in standpoints, limitations, and strengths in developing, refining, and contextualizing automated assessments across two fields. They may also disrupt the theoretical iterations of development within the broader activity system

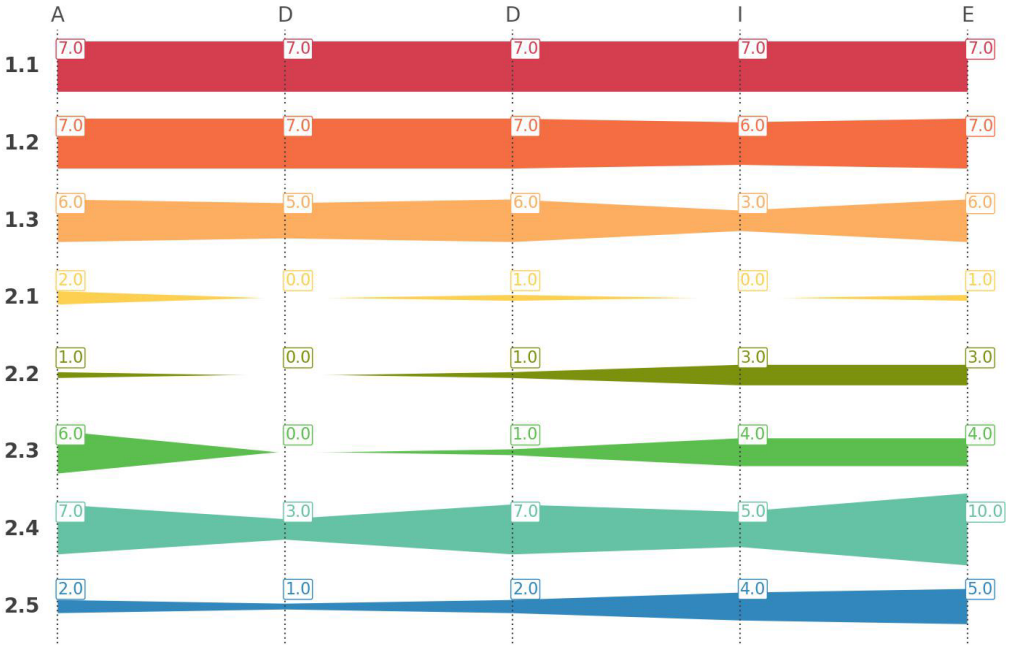

Figure 8 *Overview of action models across different motives*

Motive 1.1 = Develop automated assessment to increase pedagogical effectiveness of assessment (feedback)
Motive 1.2 = Develop automated assessment to increase evaluative function of assessment for complex tasks
Motive 1.3 = Develop automated assessment with high technical performance
Motive 2.1 = Evaluate the validity and reliability of automated assessment
Motive 2.2 = Evaluate the impact of automated assessment on learning outcomes
Motive 2.3 = Analysis and evaluate the perception, acceptance, and trust in automated assessment
Motive 2.4 = Evaluate the technical performance of automated assessment
Motive 2.5 = Evaluate the user interaction with automated assessment tool

Who are the subjects and what divisions of labor are involved in these activities?

The analysis of subject involvement reveals five main categories of, stakeholders, developers/technicians, researchers, instructors/teachers/raters, learners, and others, each participating in the ADDIE phases differently (see Figure 9).

Researchers participate in all phases of the ADDIE model across all motives due to their role in existing studies. However, there is a trend of reduced researcher involvement in the implementation phases across all activities. Researchers are actively involved in the evaluation phases to assess technical performance and may also act as developers, which is not fully represented in this research.

Developers' roles are primarily limited to the design and development phases, except for activities with motive 2.3. They are rarely involved in the implementation phases across all motives, indicating a potential disconnect between developers and end users regarding interaction with the automated assessments they create.

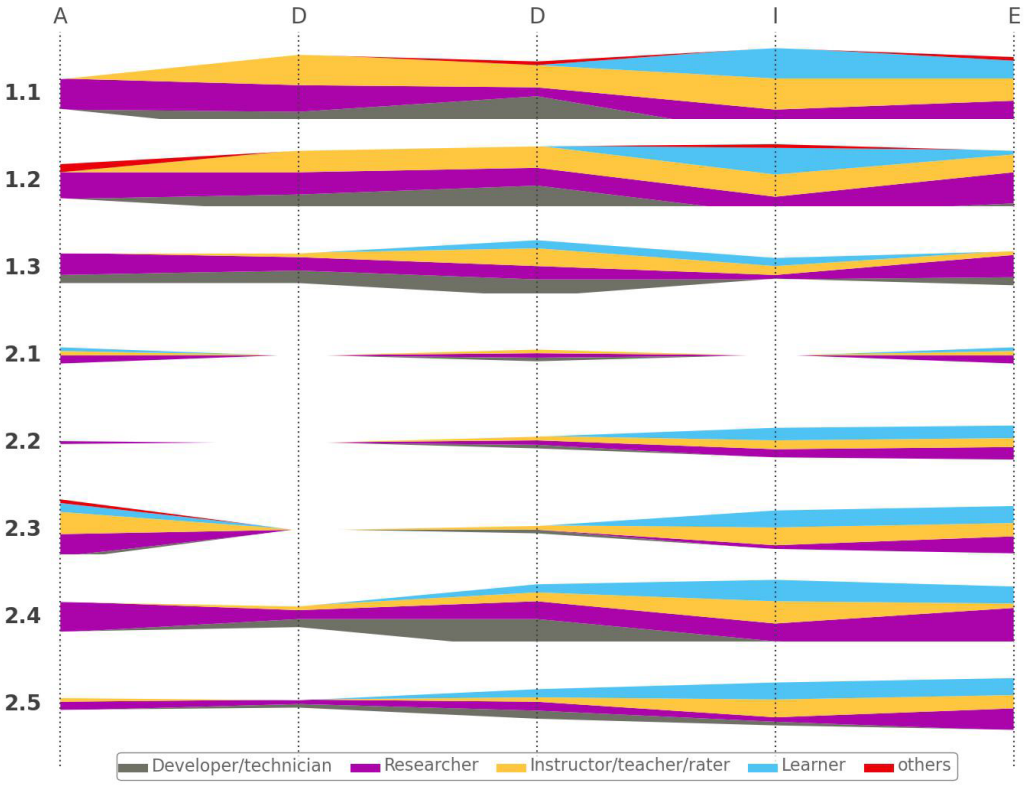

Figure 9 Overview of subject involvement in the ADDIE phases across different motives

Motive 1.1 = Develop automated assessment to increase pedagogical effectiveness of assessment (feedback)
Motive 1.2 = Develop automated assessment to increase evaluative function of assessment for complex tasks
Motive 1.3 = Develop automated assessment with high technical performance
Motive 2.1 = Evaluate the validity and reliability of automated assessment
Motive 2.2 = Evaluate the impact of automated assessment on learning outcomes
Motive 2.3 = Analysis and evaluate the perception, acceptance, and trust in automated assessment
Motive 2.4 = Evaluate the technical performance of automated assessment
Motive 2.5 = Evaluate the user interaction with automated assessment tool

Regarding the role of instructor, the findings suggest that instructors are in the activities across all motives. However, they are underrepresented in the design and evaluation phases when activities focus on technical advancement (motive 1.3 and motive 2.4), thus indicating that instructors are not actively participating in the technical design of automated assessments. Qualitative analysis indicates that while instructors have been involved in different ADDIE phases, their roles mainly involve designing assessment questions and answers questions and answers (n=7) [25], [26], [27], [32], [33], [34], [49], and setting grading criteria (n=9) [25], [29], [32], [33], [34], [35], [36], [38], [57]. Instructors primarily participate passively by providing data for training models and examining model performance. Very few studies have engaged instructors as active stakeholders, such as defining the pedagogical functionality of automated grading tools (n=3) [24], [28], [33], evaluating pedagogical alignment (n=4) [24], [26], [28], [35], and providing feedback in human-in-the-loop development to refine grading model accuracy (n=6) [24], [27], [28], [32], [46], [50].

As for students, they are almost absent from the design and development phases across all activities with different motives. They are mainly involved in the implementation and evaluation phases of activities focusing on their learning process and perceptions (motives 1.1, 2.1, 2.2, 2.3, and 2.5). Learners are seldom involved in evaluating automated grading tools for summative assessment (motive 1.2) and the technical performance of automated assessments (motive 1.3). Students' roles are mainly passive, such as generating data for model training and testing, as well as participating in learning activities involving automated grading tools. Only some activities involve students providing feedback on the user experience of automated grading tools (n=10) [7], [24], [29], [35], [44], [45], [46], [47], [51], [60], raising ethical consideration (n=2) (34, 76) and reflecting on the impact on their learning (n=12) [24], [29], [35], [44], [45], [46], [47], [50], [51], [53], [60], [61].

The “others” category has minimal involvement in all activities. Only five activities indicate contributions from various stakeholders, including a linguist developing language and rhetorical structures (motive 1.1), an external expert for ground truth grading (motive 1.1), faculty members assessing feasibility and need (motive 1.2), and stakeholders with AI experience discussing acceptance in decision-making (motive 2.3). Despite these contributions, there is a need for a better understanding of experts’ roles from different fields and their contributions to the development of automated grading tools.

DISCUSSION AND CONCLUSION

How are automated grading tools developed and integrated into educational assessment activities from an interdisciplinary perspective?

This review identifies two main activities advancing automated assessments in educational contexts (see figure 10). Activity 1 focuses on developing reliable and accurate automated assessments with educational value, while activity 2 evaluates the performance, usage, and impact of these tools. Several breakpoints are identified, which have hindered the development of ethical automated assessments that align with educational theories. These breakpoints occur in the motives of existing activity systems, the assessment purposes of the automated grading tools, the evaluation indicators used to assess the effectiveness and impact of automated assessments (i.e., rules) and the interdisciplinary collaboration involving subjects (i.e., subjects, action models, and division of labor).

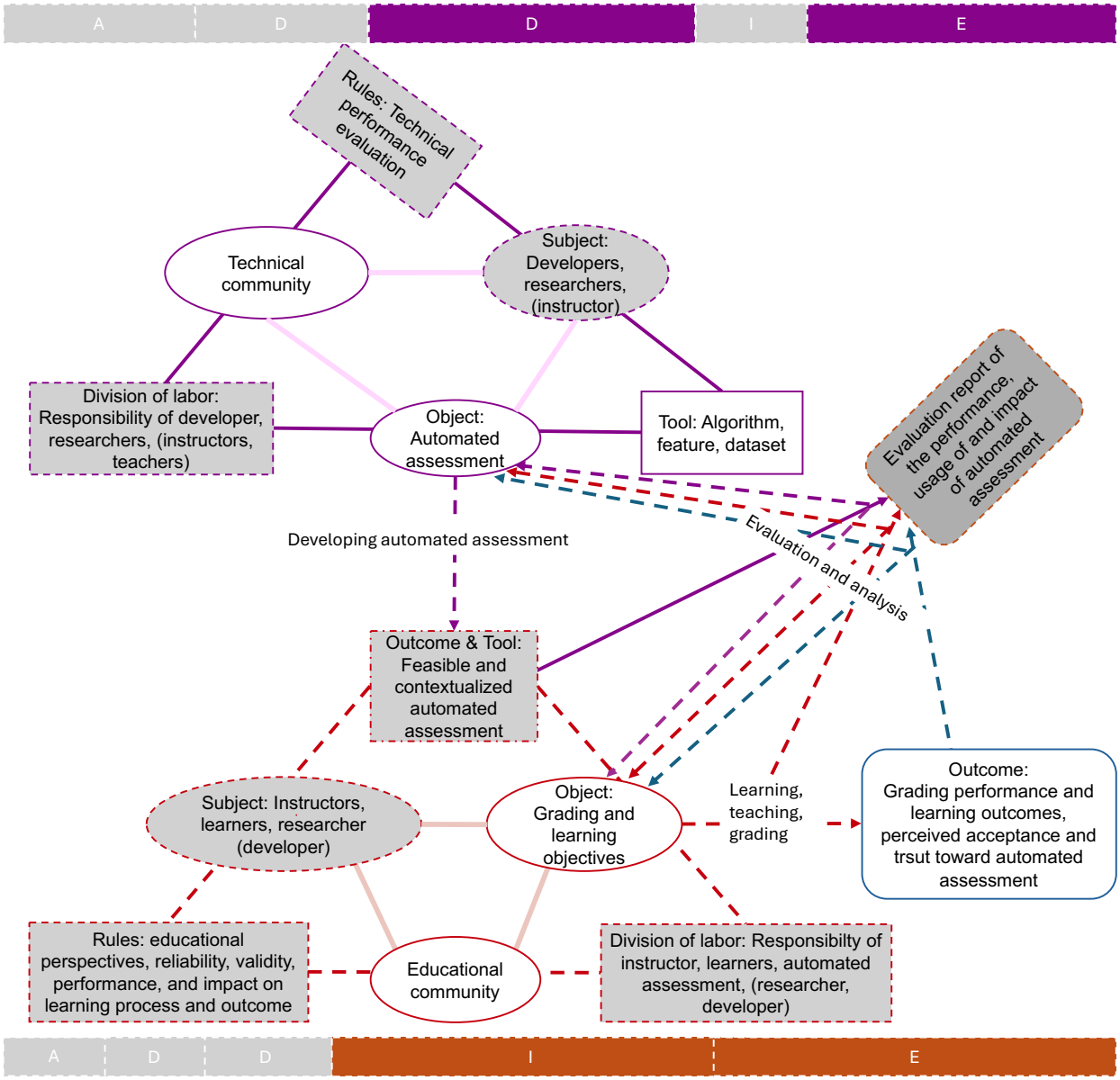

Figure 10 Activity system of advancing AI-driven automated assessments in educational contexts

he breakpoints in the motives of the existing activities highlight fundamental issues in the field. First, there is a need for understanding the validity and reliability of automated assessments from both technical and educational perspectives to determine their true performance. Despite concerns that automated grading tools fail to capture complex constructs [4], [6], efforts to address these issues are insufficient to focusing on the construct validity of algorithmic grading [13]. Second, more activities are needed to evaluate the ethical element, i.e., transparency, justice, fairness, non-maleficence, responsibility, and privacy, of the automated grading tools from both objective performance metrics and human-centric evaluations [19], [63]. Third, there is a lack of activities focusing on the interaction between the users, such as instructors and teachers, and automated assessments, and its impact on learning outcomes. This gap hinders the comprehension of how users might constructively or otherwise use automated assessments, affecting long-term learning outcomes that are not immediately observable.

The breakpoints in the assessment purposes of automated grading tools, particularly technical advancement, lies in their emphasis on summative assessment. This lack of technical advancement of automated grading tools for formative assessments suggest that the system might rely on summative feedback mechanisms without proper alignment with learning interactions [4]. Another critical break point is the lack of deliberate actions to use automated assessment as a learning tool to develop learners' constructive role in understanding, evaluating, and using automated assessments for self-regulation [64]. Further research is necessary to understand the impact of learners' AI literacy and critical thinking on automated assessment development and their role in fostering self-regulated learning and contributing to tool development.

Breakpoints also exist in the indicators used to evaluate the performance and impact of automated assessments. First, there is an urgent need to research the construct validity of automated grading tools instead of over-focusing on the human-machine agreement. Secondly, there is an overemphasis on technical performance, neglecting ethical considerations and the objective impact on learning and teaching from both technical and educational perspectives. Thirdly, the differing evaluation criteria and indicators used by technical and educational perspectives highlight gaps in mutual understanding, necessitating further research to achieve synergy. It is crucial to understand how developers, researchers, instructors, and learners can play constructive and proactive roles in addressing these gaps. For example, while technical definitions of reliability, validity, and accuracy emphasize probabilistic results, instructors and students need to interpret these results effectively to avoid an all-or-nothing manner. Conversely, developers should enhance the transparency and explainability of automated grading tools to support teachers' and students' activities [63].

Breakpoints in action models and divisions of labor reveal insufficient active engagement of instructors and learners in the development and use of these tools, missing their constructive and proactive roles. In the action model, technical efforts prioritize development and performance evaluation but often neglect real-world implementation, while educational efforts focus on implementation and impact, omitting model refinement. Limited collaborative efforts are reported in the alignment of automated grading tools with pedagogical requirements, defining assessment criteria, converting pedagogical requirements to technical specifications, ensuring data reliability, and defining key features [1]. These collaborations are essential for enhancing the validity, reliability, transparency, and explainability of automated assessments. Additionally, students' constructive roles in using automated assessments for improving their learning are under researched, indicating the need of investigating and developing their AI literacy and cognitive competencies for using automated grading tools. Moreover, professional development from the perspectives of instructors is necessary.

This review underscores the complexity and interconnectedness of developing and implementing automated assessment tools in educational contexts. While technical advancements are valuable, it's crucial to contextualize these tools and consider their practical educational applications. Beyond striving for human-machine agreement, we should scrutinize the construct validity these tools introduce, including their ability to assess constructs that may suppass human evaluation. Discussions should focus on how these tools can complement human limitations and work synergistically with educators, rather than merely replacing or dismissing human evaluators. Engaging developers, instructors, and learners as active, agentic and constructive participants is essential. By understanding the impact of automatic grading tools on teaching and learning, stakeholders can enhance these technologies through feedback. As AI increasingly influences society, it's vital for instructors and students to

comprehend both the outcomes generated by automatic grading systems and the mechanisms behind them.

Note: OpenAI's ChatGPT-4 has been used to review language grammar and readability for this manuscript.

REFERENCE

[1] H. Crompton and D. Burke, "Artificial intelligence in higher education: the state of the field," *International Journal of Educational Technology in Higher Education*, vol. 20, no. 1, p. 22, Apr. 2023, doi: 10.1186/s41239-023-00392-8.

[2] R. Weegar and P. Idestam-Almquist, "Reducing Workload in Short Answer Grading Using Machine Learning," *Int J Artif Intell Educ*, Feb. 2023, doi: 10.1007/s40593-022-00322-1.

[3] E. Dimitriadou and A. Lanitis, "A critical evaluation, challenges, and future perspectives of using artificial intelligence and emerging technologies in smart classrooms," *Smart Learning Environments*, vol. 10, no. 1, p. 12, Feb. 2023, doi: 10.1186/s40561-023-00231-3.

[4] J. Gardner, M. O'Leary, and L. Yuan, "Artificial intelligence in educational assessment: 'Breakthrough? Or buncombe and ballyhoo?,'" *Journal of Computer Assisted Learning*, vol. 37, no. 5, pp. 1207–1216, 2021, doi: 10.1111/jcal.12577.

[5] K. Ercikan and D. F. McCaffrey, "Optimizing Implementation of Artificial-Intelligence-Based Automated Scoring: An Evidence Centered Design Approach for Designing Assessments for AI-based Scoring," *Journal of Educational Measurement*, vol. 59, no. 3, pp. 272–287, 2022, doi: 10.1111/jedm.12332.

[6] D. Ramesh and S. K. Sanampudi, "An automated essay scoring systems: A systematic literature review," *Artificial Intelligence Review*, vol. 55, no. 3, pp. 2495–2527, 2022, [Online]. Available: dadiramesh44@gmail.com, sureshsanampudi@jntuh.ac.in, ORCID: 0000-0002-3967-8914

[7] C. M. Barrett, "Automated essay evaluation and the computational paradigm: Machine scoring enters the classroom," Ph.D., University of Rhode Island, United States -- Rhode Island, 2015. [Online]. Available: https://www.proquest.com/dissertations-theses/automated-essay-evaluation-computational-paradigm/docview/1710482685/se-2?accountid=14774

[8] S. Jackson and N. Panteli, "Trust or mistrust in algorithmic grading? An embedded agency perspective.," *International Journal of Information Management*, vol. 69, p. N.PAG-N.PAG, 2023, [Online]. Available: https://search.ebscohost.com/login.aspx?direct=true&db=lxh&AN=161814539&site=ehost-live

[9] A. Horbach and M. Pinkal, "Semi-Supervised Clustering for Short Answer Scoring," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, and T. Tokunaga, Eds., Miyazaki, Japan: European Language Resources Association (ELRA), May 2018. Accessed: Jul. 10, 2024. [Online]. Available: https://aclanthology.org/L18-1641

[10] F. Zehner, C. Sälzer, and F. Goldhammer, "Automatic coding of short text responses via clustering in educational assessment," *Educational and Psychological Measurement*, vol. 76, no. 2, pp. 280–303, 2016, [Online]. Available:

fabian.zehner@tum.de, ORCID: 0000-0003-3512-1403, ORCID: 0000-0003-0289-9534

[11] B. Beigman Klebanov and N. Madnani, “Automated Evaluation of Writing – 50 Years and Counting,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online: Association for Computational Linguistics, Jul. 2020, pp. 7796–7810. doi: 10.18653/v1/2020.acl-main.697.

[12] R. Bennett, “Moving the Field Forward: Some Thoughts on Validity and Automated Scoring,” Jan. 2004.

[13] R. J. Mislevy, R. G. Almond, and J. F. Lukas, “A Brief Introduction to Evidence-Centered Design,” *ETS Research Report Series*, vol. 2003, no. 1, pp. i–29, 2003, doi: 10.1002/j.2333-8504.2003.tb01908.x.

[14] R. Dann, “Assessment *as* learning: blurring the boundaries of assessment and learning for theory, policy and practice,” *Assessment in Education: Principles, Policy & Practice*, vol. 21, no. 2, pp. 149–166, Apr. 2014, doi: 10.1080/0969594X.2014.898128.

[15] L. H. Schellekens, H. G. J. Bok, L. H. de Jong, M. F. van der Schaaf, W. D. J. Kremer, and C. P. M. van der Vleuten, “A scoping review on the notions of Assessment as Learning (AaL), Assessment for Learning (AfL), and Assessment of Learning (AoL),” *Studies in Educational Evaluation*, vol. 71, p. 101094, Dec. 2021, doi: 10.1016/j.stueduc.2021.101094.

[16] D. Boud and N. Falchikov, “Aligning Assessment with Long-Term Learning,” *Assessment & Evaluation in Higher Education - ASSESS EVAL HIGH EDUC*, vol. 31, pp. 399–413, Aug. 2006, doi: 10.1080/02602930600679050.

[17] S. Burrows, I. Gurevych, and B. Stein, “The Eras and Trends of Automatic Short Answer Grading,” *International Journal of Artificial Intelligence in Education*, vol. 25, no. 1, pp. 60–117, Mar. 2015, [Online]. Available: http://dx.doi.org/10.1007/s40593-014-0026-8

[18] S. Valenti, F. Neri, and A. Cucchiarelli, “An Overview of Current Research on Automated Essay Grading,” *Journal of Information Technology Education. Research*, vol. 2, pp. 319–330, 2003, doi: 10.28945/331.

[19] A. Jobin, M. Ienca, and E. Vayena, “The global landscape of AI ethics guidelines,” *Nat Mach Intell*, vol. 1, no. 9, pp. 389–399, Sep. 2019, doi: 10.1038/s42256-019-0088-2.

[20] Y. Engeström, “Learning in Activity,” in *The Cambridge Handbook of the Learning Sciences*, 3rd ed., R. K. Sawyer, Ed., Cambridge University Press, 2022, pp. 134–155. doi: 10.1017/9781108888295.009.

[21] Y. Engeström, R. Miettinen, and R.-L. Punamäki-Gitai, *Perspectives on Activity Theory*. Cambridge University Press, 1999.

[22] K. Kuurti, “Activity Theory as a Potential Framework for Human-Computer Interaction Research,” in *Context and Consciousness*, B. A. Nardi, Ed., The MIT Press, 1995, pp. 17–44. doi: 10.7551/mitpress/2137.003.0006.

[23] M. J. Page *et al.*, “PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews,” *BMJ*, p. n160, Mar. 2021, doi: 10.1136/bmj.n160.

[24] S. Knight, S. Buckingham Shum, P. Ryan, Á. Sándor, and X. Wang, “Designing academic writing analytics for civil law student self-assessment,” *International Journal of Artificial Intelligence in Education*, vol. 28, no. 1, pp. 1–28, 2018, [Online].

Available: sjgknight@gmail.com, first.lastname@uts.edu.au, agnes.sandor@xrce.xerox.com, ORCID: 0000-0002-8709-5780
[25] A. Kurnia, A. Lim, and B. Cheang, “Online judge.,” *Computers & Education*, vol. 36, no. 4, pp. 299–315, 2001, [Online]. Available: https://search.ebscohost.com/login.aspx?direct=true&db=eue&AN=507773322&site=ehost-live
[26] O. H. T. Lu, A. Y. Q. Huang, D. C. L. Tsai, and S. J. H. Yang, “Expert-Authored and Machine-Generated Short-Answer Questions for Assessing Students’ Learning Performance.,” *Journal of Educational Technology & Society*, vol. 24, no. 3, pp. 159–173, 2021, [Online]. Available: https://search.ebscohost.com/login.aspx?direct=true&db=eue&AN=151577181&site=ehost-live
[27] A. Malik *et al.*, “Generative Grading: Near Human-Level Accuracy for Automated Feedback on Richly Structured Problems.” International Educational Data Mining Society, Jan. 01, 2021. [Online]. Available: https://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=ED615516&site=ehost-live
[28] K. Matthews, T. Janicki, L. He, and L. Patterson, “Implementation of an Automated Grading System with an Adaptive Learning Component to Affect Student Feedback and Response Time.,” *Journal of Information Systems Education*, vol. 23, no. 1, pp. 71–83, 2012, [Online]. Available: https://search.ebscohost.com/login.aspx?direct=true&db=eue&AN=79352981&site=ehost-live
[29] F. Rodrigues and P. Oliveira, “A system for formative assessment and monitoring of students’ progress,” *Computers & Education*, vol. 76, pp. 30–41, Jul. 2014, doi: 10.1016/j.compedu.2014.03.001.
[30] P. Vittorini, S. Menini, and S. Tonelli, “An AI-Based System for Formative and Summative Assessment in Data Science Courses.,” *International Journal of Artificial Intelligence in Education (Springer Science & Business Media B.V.)*, vol. 31, no. 2, pp. 159–185, 2021, [Online]. Available: https://search.ebscohost.com/login.aspx?direct=true&db=eue&AN=150768309&site=ehost-live
[31] C. Geigle, C. Zhai, and D. C. Ferguson, “An Exploration of Automated Grading of Complex Assignments,” in *Proceedings of the Third (2016) ACM Conference on Learning @ Scale*, Edinburgh Scotland UK: ACM, Apr. 2016, pp. 351–360. doi: 10.1145/2876034.2876049.
[32] E. E. Maicus, “Automated Grading for Advanced Topics Courses,” Ph.D., Rensselaer Polytechnic Institute, United States -- New York, 2021. [Online]. Available: https://www.proquest.com/dissertations-theses/automated-grading-advanced-topics-courses/docview/2729068668/se-2?accountid=14774
[33] F. H. Psalmerosi, “Applying Text Mining and Machine Learning to Build Methods for Automated Grading.” Accessed: Sep. 03, 2024. [Online]. Available: https://essay.utwente.nl/77190/
[34] L. D. Ried PhD *et al.*, “An Automated Competency-based Student Performance Assessment Program for Advanced Pharmacy Practice Experiential Programs,” *American Journal of Pharmaceutical Education*, vol. 71, no. 6, pp. 1–128, 2007, [Online]. Available: https://www.proquest.com/scholarly-journals/automated-

competency-based-student-performance/docview/211257060/se-2?accountid=14774
[35] B. Tomic, A. Kijevcanin, Z. Sevarac, and J. M. Jovanovic, “An AI-based Approach for Grading Students’ Collaboration,” *IEEE Transactions on Learning Technologies*, vol. 16, no. 3, pp. 292–305, 2023, doi: 10.1109/TLT.2022.3225432.
[36] H.-C. Wang, C.-Y. Chang, and T.-Y. Li, “Assessing creative problem-solving with automated text grading.,” *Computers & Education*, vol. 51, no. 4, pp. 1450–1466, 2008, [Online]. Available: https://search.ebscohost.com/login.aspx?direct=true&db=eue&AN=508020508&site=ehost-live
[37] V. S. Kumar and D. Boulanger, “Automated Essay Scoring and the Deep Learning Black Box: How Are Rubric Scores Determined?,” *International Journal of Artificial Intelligence in Education (Springer Science & Business Media B.V.)*, vol. 31, no. 3, pp. 538–584, 2021, [Online]. Available: https://search.ebscohost.com/login.aspx?direct=true&db=eue&AN=152372442&site=ehost-live
[38] A. V. Y. Lee, A. C. Luco, and S. C. Tan, “A Human-Centric Automated Essay Scoring and Feedback System for the Development of Ethical Reasoning,” 2024.
[39] D. E. Powers, D. S. Escoffery, and M. P. Duchnowski, “Validating Automated Essay Scoring: A (Modest) Refinement of the ‘Gold Standard,’” *Applied Measurement in Education*, vol. 28, no. 2, p. 130, 2015, [Online]. Available: https://www.proquest.com/scholarly-journals/validating-automated-essay-scoring-modest/docview/1671095320/se-2?accountid=14774
[40] S. K. Saha and R. C. Dhawaleswar, “Development of a practical system for computerized evaluation of descriptive answers of middle school level students,” *Interactive Learning Environments*, vol. 30, no. 2, pp. 215–228, Mar. 2022, doi: 10.1080/10494820.2019.1651743.
[41] M. Uto and M. Okano, “Learning Automated Essay Scoring Models Using Item-Response-Theory-Based Scores to Decrease Effects of Rater Biases,” *IEEE Transactions on Learning Technologies*, vol. 14, no. 6, pp. 763–776, 2021, doi: 10.1109/TLT.2022.3145352.
[42] T. K. Landauer, “Automatic Essay Assessment,” *Assessment in Education: Principles, Policy & Practice*, vol. 10, no. 3, pp. 295–308, Nov. 2003, doi: 10.1080/0969594032000148154.
[43] Ö. G. Ulum, “A critical deconstruction of computer-based test application in Turkish State University,” *Education and Information Technologies*, vol. 25, no. 6, pp. 4883–4896, Nov. 2020, doi: 10.1007/s10639-020-10199-z.
[44] J. Choi, “The impact of Automated Essay Scoring (AES) for improving English language learner’s essay writing,” Ph.D., University of Virginia, United States -- Virginia, 2010. [Online]. Available: https://www.proquest.com/dissertations-theses/impact-automated-essay-scoring-aes-improving/docview/821699679/se-2?accountid=14774
[45] X. Lv, “A Study on the Application of Automatic Scoring and Feedback System in College English Writing,” *International Journal of Emerging Technologies in Learning (Online)*, vol. 13, no. 3, pp. 188–196, 2018, doi: 10.3991/ijet.v13i03.8386.
[46] S. H. S. Tan, G. Thibault, A. C. Y. Chew, and P. Rajalingam, “Enabling open-ended questions in team-based learning using automated marking: Impact on student

achievement, learning and engagement," *Journal of Computer Assisted Learning*, vol. 38, no. 5, pp. 1347–1359, 2022, doi: 10.1111/jcal.12680.
[47] G. J. Coulthard, "A descriptive case study: Investigating the implementation of web based, automated grading and tutorial software in a freshman computer literacy course," Ph.D., Purdue University, United States -- Indiana, 2016. [Online]. Available: https://www.proquest.com/dissertations-theses/descriptive-case-study-investigating/docview/1875556393/se-2?accountid=14774
[48] C. Dreher, T. Reiners, and H. Dreher, "Investigating Factors Affecting the Uptake of Automated Assessment Technology," *Journal of Information Technology Education*, vol. 10, pp. 161–181, Jan. 2011, [Online]. Available: http://www.jite.org/documents/Vol10/JITEv10p161-181Dreher950.pdf
[49] C. Greiner, T. C. Peisl, F. Höpfl, and O. Beese, "Acceptance of AI in Semi-Structured Decision-Making Situations Applying the Four-Sides Model of Communication—An Empirical Analysis Focused on Higher Education," *Education Sciences*, vol. 13, no. 9, p. 865, 2023, doi: 10.3390/educsci13090865.
[50] S. Hsu, T. W. Li, Z. Zhang, M. Fowler, C. Zilles, and K. Karahalios, "Attitudes Surrounding an Imperfect AI Autograder," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, in CHI '21. New York, NY, USA: Association for Computing Machinery, May 2021, pp. 1–15. doi: 10.1145/3411764.3445424.
[51] Y. Lai, "Which Do Students Prefer to Evaluate Their Essays: Peers or Computer Program," *British Journal of Educational Technology*, vol. 41, no. 3, pp. 432–454, May 2010, [Online]. Available: http://dx.doi.org/10.1111/j.1467-8535.2009.00959.x
[52] S. A. Nagro, "The Role of Artificial Intelligence Techniqies in Improving the Behavior and Practices of Faculty Members When Switching to Elearning in Light of the COVID-19 Crisis," *International Journal of Education and Practice*, vol. 9, no. 4, pp. 687–714, Jan. 2021, [Online]. Available: https://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=EJ1329070&site=ehost-live
[53] T. Reiners, C. Dreher, and H. Dreher, "Six Key Topics for Automated Assessment Utilisation and Acceptance," *Informatics in Education*, vol. 10, no. 1, pp. 47–64, Jan. 2011, [Online]. Available: https://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=EJ1064287&site=ehost-live
[54] N. Andersen and F. Zehner, "shinyReCoR: A Shiny Application for Automatically Coding Text Responses Using R," *Psych*, vol. 3, no. 3, pp. 422–446, Aug. 2021, doi: 10.3390/psych3030030.
[55] S. Azad, B. Chen, M. Fowler, M. West, and C. Zilles, "Strategies for Deploying Unreliable AI Graders in High-Transparency High-Stakes Exams," in *Artificial Intelligence in Education*, I. I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, and E. Millán, Eds., Cham: Springer International Publishing, 2020, pp. 16–28. doi: 10.1007/978-3-030-52237-7_2.
[56] J. A. Erickson, A. F. Botelho, and Z. Peng, "Is It Fair? Automated Open Response Grading".
[57] O. L. Liu, C. Brew, J. Blackmore, L. Gerard, J. Madhok, and M. C. Linn, "Automated Scoring of Constructed-Response Science Items: Prospects and Obstacles," *Educational Measurement, Issues and Practice*, vol. 33, no. 2, p. 19, Summer 2014,

[Online]. Available: https://www.proquest.com/scholarly-journals/automated-scoring-constructed-response-science/docview/1543777355/se-2?accountid=14774
[58] E. D. Reilly, Rose Eleanore Stafford, K. M. Williams, and Stephanie Brooks Corliss, "Evaluating the validity and applicability of automated essay scoring in two massive open online courses," *International Review of Research in Open and Distance Learning, suppl. Special Issue: Research into Massive Open Online Courses*, vol. 15, no. 5, Nov. 2014, [Online]. Available: https://www.proquest.com/scholarly-journals/evaluating-validity-applicability-automated-essay/docview/1634290940/se-2?accountid=14774
[59] J. Schneider, R. Richner, and M. Riser, "Towards Trustworthy AutoGrading of Short, Multi-lingual, Multi-type Answers.," *International Journal of Artificial Intelligence in Education (Springer Science & Business Media B.V.)*, vol. 33, no. 1, pp. 88–118, 2023, [Online]. Available: https://search.ebscohost.com/login.aspx?direct=true&db=eue&AN=162057551&site=ehost-live
[60] M. Tsai, "The Consistency Between Human Raters and an Automated Essay Scoring System in Grading High School Students' English Writing.," *Action in Teacher Education (Association of Teacher Educators)*, vol. 34, no. 4, pp. 328–335, 2012, [Online]. Available: https://search.ebscohost.com/login.aspx?direct=true&db=eue&AN=83770340&site=ehost-live
[61] M.-H. Tsai, "The Most Preferred and Effective Reviewer of L2 Writing among Automated Grading System, Peer Reviewer and Teacher," *World Journal of Education*, vol. 7, no. 4, pp. 60–84, Jan. 2017, [Online]. Available: https://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=EJ1157404&site=ehost-live
[62] J. Liebenberg and V. Pieterse, "Investigating the Feasibility of Automatic Assessment of Programming Tasks," *Journal of Information Technology Education: Innovations in Practice*, vol. 17, pp. 201–223, Jan. 2018, [Online]. Available: http://www.jite.org/documents/Vol17/JITEv17IIPp201-223Liebenberg4853.pdf
[63] G. Vilone and L. Longo, "Notions of explainability and evaluation approaches for explainable artificial intelligence," *Information Fusion*, vol. 76, pp. 89–106, Dec. 2021, doi: 10.1016/j.inffus.2021.05.009.
[64] R. Conijn, R. Martinez-Maldonado, S. Knight, S. Buckingham Shum, L. Van Waes, and M. Van Zaanen, "How to provide automated feedback on the writing process? A participatory approach to design writing analytics tools," *Computer Assisted Language Learning*, vol. 35, no. 8, pp. 1838–1868, Nov. 2022, doi: 10.1080/09588221.2020.1839503.