ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/381769977

The Rise of Artificial Intelligence in Educational Measurement: Opportunities and Ethical Challenges

Preprint · June 2024

DOI: 10.48550/arXiv.2406.18900

CITATIONS
0

READS
523

17 authors, including:

Okan Bulut
University of Alberta
188 PUBLICATIONS **2,074** CITATIONS
SEE PROFILE

Jodi M Casabianca
Educational Testing Service
26 PUBLICATIONS **488** CITATIONS
SEE PROFILE

Hong Jiao
University of Maryland, College Park
96 PUBLICATIONS **1,327** CITATIONS
SEE PROFILE

Joshua Wilson
University of Delaware
51 PUBLICATIONS **1,332** CITATIONS
SEE PROFILE

All content following this page was uploaded by Okan Bulut on 28 June 2024.

The user has requested enhancement of the downloaded file.

arXiv:2406.18900v1 [cs.CY] 27 Jun 2024

The Rise of Artificial Intelligence in Educational Measurement: Opportunities and Ethical Challenges

Okan Bulut[1,*], Maggie Beiting-Parrish[2], Jodi M. Casabianca[3], Sharon C. Slater[3], Hong Jiao[4], Dan Song[5], Christopher Ormerod[6], Deborah Gbemisola Fabiyi[7], Rodica Ivan[8], Cole Walsh[8], Oscar Rios[9], Joshua Wilson[10], Seyma N. Yildirim-Erbasli[11], Tarid Wongvorachan[1], Joyce Xinle Liu[1], Bin Tan[1], Polina Morilova[1]

[1] University of Alberta

[2] Federation of American Scientists

[3] Educational Testing Service

[4] University of Maryland

[5] University of Iowa

[6] Cambium Assessment

[7] Washington State University

[8] Acuity Insights

[9] PSI Services

[10] University of Delaware

[11] Concordia University of Edmonton

These authors contributed equally to this work

∗ Corresponding author; bulut@ualberta.ca

June 2024

Acknowledgement

We extend our sincere gratitude to the Special Interest Group on Artificial Intelligence in Measurement and Education (AIME) and the National Council of Measurement in Education (NCME) for initiating this white paper.

Abstract

The integration of artificial intelligence (AI) in educational measurement has revolutionized assessment methods, enabling automated scoring, rapid content analysis, and personalized feedback through machine learning and natural language processing. These advancements provide timely, consistent feedback and valuable insights into student performance, thereby enhancing the assessment experience. However, the deployment of AI in education also raises significant ethical concerns regarding validity, reliability, transparency, fairness, and equity. Issues such as algorithmic bias and the opacity of AI decision-making processes pose risks of perpetuating inequalities and affecting assessment outcomes. Responding to these concerns, various stakeholders, including educators, policymakers, and organizations, have developed guidelines to ensure ethical AI use in education. The National Council of Measurement in Education's Special Interest Group on AI in Measurement and Education (AIME) also focuses on establishing ethical standards and advancing research in this area. In this paper, a diverse group of AIME members examines the ethical implications of AI-powered tools in educational measurement, explores significant challenges such as automation bias and environmental impact, and proposes solutions to ensure AI's responsible and effective use in education.

Keywords— artificial intelligence, educational measurement, LLM, bias, fairness

1 Introduction

Emerging technologies and applications powered by artificial intelligence (AI) continue to bring significant changes to every scientific field, including educational measurement. The integration of AI into educational measurement has significantly transformed the different methods that are used in practice. For example, using machine learning and deep learning algorithms, AI enables automated scoring (also referred to as automated essay scoring) that can evaluate open-ended responses, essays, and even creative work, providing faster and more consistent feedback to students. This application saves educators time while yielding immediate feedback for learners, allowing for an efficient learning experience. Furthermore, natural language processing (NLP) algorithms can be used to analyze written (or spoken) content rapidly, identify improvement areas, and offer learners personalized feedback. Beyond a conventional assessment setting, AI-driven data analytics tools can also help educators and administrators gather insights into student performance, identify trends, predict future academic outcomes, and recommend interventions to support struggling students.

Although AI-powered innovations have created promising opportunities for more robust, efficient, and personalized assessments, they have also raised significant concerns regarding validity, reliability, transparency, fairness, equity, and test security (Hao et al., 2024). For example, advanced AI algorithms (e.g., deep learning) often operate as "black boxes," making understanding how they arrive at specific decisions challenging. Such algorithms can inadvertently perpetuate or amplify biases present in the data used to train them. In educational contexts, algorithmic bias can affect assessment outcomes (e.g., test scores or grades), exacerbate existing inequalities, and disadvantage certain groups of students. Therefore, the uncontrolled and unregulated development, deployment, and utilization of AI tools in enhancing educational outcomes for learners may lead to unintended consequences, jeopardizing the effectiveness of AI-driven recommendations or assessments.

Ethical concerns have motivated different stakeholders, including educators, researchers, assessment specialists, and policymakers, to regulate the ethical use of AI in educational measurement. Addressing major challenges such as test bias has been a long-standing goal in educational measurement. However, with the rapid development of AI-powered tools in education, resolving ethical challenges has become an urgent priority (J. Zhou et al., 2020). Various government agencies, non-profit research organizations, and other institutions have developed an active agenda to establish standards of ethical use of AI in education. For example, the European Commission published ethical guidelines on the use of AI and data in teaching and learning for educators (European Commission, Directorate-General for Education, Youth & Culture, 2022). The guidelines draw attention to assessment-related topics, such as scoring short-answer items and essays using automated tools, automatic feedback on writing tasks, and algorithms for personalized assessment tools. The Organisation for Economic Co-operation and Development (OECD) also shared a policy brief that discusses the need for explainability and transparency when using digital tools powered by advanced AI technologies in high-stakes settings for students, teachers, or educational establishments (OECD, 2023). A recent endeavor that prioritized the ethical utilization of AI for assessment purposes is the implementation of the Duolingo English Test's Responsible AI Standards (Burstein et al., 2023). These standards aim to guide stakeholders on the collaborative and judicious application of AI and human expertise to ensure reliable, secure, and effective assessments.

Recently, the National Council of Measurement in Education has established the Special Interest Group on AI in Measurement and Education (AIME) to advance both theoretical and applied research on the use of AI in educational measurement. AIME consists of a diverse group of members, including data scientists, psychometricians, educational researchers, and other key stakeholders in education. One of AIME's primary objectives is to establish guidelines for the ethical use of AI in educational measurement. In this paper, we aim to contribute to this objective by examining the ethical use of AI-powered applications across various subdomains of educational measurement, such as item generation, scoring, proctoring, and feedback. Also, we discuss other ethical concerns, such as automation bias and the environmental impacts of AI tools in education. Through a detailed exploration of these subdomains, we intend to identify prevailing challenges, highlight ethical issues, and propose viable solutions where appropriate. Considering the rapid evolution of AI algorithms and systems, we anticipate this paper will be an important starting point for researchers and practitioners working on educational measurement, serving as a strong foundation for future research into the new ethical challenges that may emerge from AI-powered assessment tools.

2 Automated Item Generation

As the testing industry transitioned from paper-and-pencil to digital formats over the last two decades, the demand for large quantities of high-quality items has increased substantially. Digital assessments have revolutionized how tests are administered and results are analyzed, offering greater flexibility and precision. For example, computerized adaptive testing (CAT) dynamically adjusts the difficulty of questions based on the test-taker's performance in real time. If a student answers a question correctly, the next question is more challenging; if the answer is incorrect, the next question is easier. This approach aims to maintain an optimal difficulty level personalized for each examinee, thereby providing a more accurate measure of their abilities (Gorgun & Bulut, 2023; Weiss & Kingsbury, 1984). Another example is multistage adaptive testing (MST), which divides the test into several stages, each consisting of a set of items. Based on the examinee's performance in the initial stage, the system selects the most appropriate set of items for the subsequent stages. This method balances the precision of ability estimation with practical considerations, such as test length and item exposure (Bulut, Gorgun, & Karamese, 2023; Zenisky et al., 2009).

Effective item selection in these digital formats requires a large number of high-quality items to control the item exposure rate while finding optimal items from an adequately sized item pool. For instance, adaptive testing programs, such as the Graduate Record Examinations (GRE) and the National Council Licensure Examination (NCLEX), maintain extensive item banks, ensuring that each examinee receives a unique set of questions tailored to their ability level, while also safeguarding the integrity and security of the test content. This process underscores the critical need for a robust item pool to support the sophisticated algorithms that drive CAT and MST systems, ultimately enhancing the accuracy and fairness of digital assessments.

In response to this growing demand, automatic item generation (AIG) was originally proposed over a decade ago by educational measurement researchers as a cost-effective solution to generate a large number of high-quality items (e.g., Gierl & Haladyna, 2012; Holling et al., 2009; Lai et al., 2009). The most common method of AIG was to use computer algorithms to generate items based on cognitive models and item models developed by human experts, such as subject matter experts, followed by the evaluation of items by experts (Gierl & Lai, 2012). Although this method has proven effective in generating a large number of high-quality items, it still heavily relies on human input (i.e., subject matter experts) during the item generation process and yields similar items that may not be diverse enough in terms of content and item structure.

Emerging technologies, such as generative AI and large language models (LLMs), have led researchers to explore their usefulness as an alternative method for item generation. For example, Offerijns et al. (2020) used OpenAI's GPT-2 to generate many items with contextual paragraphs and answers as input text. They also used Devlin et al. (2019)'s Bidirectional Encoder Representations from Transformers (BERT) model to filter out items that were not answerable or not coherent. Kumari et al. (2022) used Google's Text-to-Text Transfer Transformer (T5) to detect answers from source texts, and then they combined the source texts and answers to generate items. In another study, Bulut and Yildirim-Erbasli (2022) used OpenAI's GPT-2 for text generation and Google's T5 for item generation associated with the generated text. In a more recent study, Jiao et al. (2023) compared the model performance of three LLMs for the AIG task by considering the coherence and creativity of the automatically generated items.

A recent scoping review summarized the existing work of leveraging LLMs for AIG (Tan et al., 2024). The review identified the commonly used LLMs and their specific usages in the AIG process, as well as the characteristics of the generated items. It concluded that LLMs are a flexible and effective solution for generating various types of items across different languages and subject domains. Although the review suggests that leveraging generative AI tools (i.e., LLMs) is a promising solution for AIG, it also revealed that many of the existing AIG studies lack a solid educational foundation. These findings underscore the need to align item generation with assessment purposes and to integrate measurement and learning theories into the AIG process.

According to Tan et al. (2024), from both practical and ethical standpoints, AIG should not conclude with merely generating a large number of items but rather with ensuring that the generated items are of high quality for use in educational contexts. However, many existing AIG studies did not involve empirical testing to evaluate the measurement properties of the generated items, unlike traditional item development studies. The measurement properties encompass a variety of item-level or test-level attributes such as item parameters, reliability, validity, and fairness. For instance, it is crucial for a test to include items with varying difficulty levels to accurately measure students' abilities with minimal measurement errors. Neglecting these critical measurement properties could lead to serious ethical issues, potentially resulting in erroneous conclusions about students' abilities, which could unfairly influence their educational paths. For example, in low-stakes assessments, such as practice tests, students might not benefit if they are not properly designed to identify their misconceptions and provide valuable feedback about their learning progress. Therefore, test developers must integrate rigorous pre-testing and validation protocols within the AIG framework. Encouragingly, this gap has been swiftly addressed as more and more researchers have realized the importance of evaluating measurement properties after item generation in AIG (e.g., Säuberli & Clematide, 2024).

AIG research should adopt a human-in-the-loop framework—a collaborative approach to integrate human expertise into AI-based decision-making. Relying solely on computer-generated items raises ethical concerns, given that assessments play a fundamental role in evaluating students' learning outcomes and monitoring the performance of larger entities such as schools and education systems (Gierl et al., 2022; Sayin & Gierl, 2024). These assessments can potentially inform educational policies and influence decisions that affect students' lives, such as their education and career paths. Consider a scenario where automated tests contain errors or fail to measure students' abilities accurately; accountability becomes ambiguous. Thus, human oversight is crucial to uphold the reliability of assessments by comprehensively monitoring and supervising AIG systems. Accordingly, we advocate for multidisciplinary teams involving subject matter experts, educators, measurement specialists, and NLP researchers to increase AI's accountability and better harness the utility and potential of LLMs for AIG. For example, NLP researchers can advance the technical aspects of AIG, ensuring the best use of LLMs according to their characteristics and features. Subject matter experts can contribute by providing their expertise and helping to create assessment items that are academically rigorous and aligned with curriculum standards. Measurement specialists can evaluate the functionality of the assessment items (e.g., reliability, validity, and fairness), examining how well they gauge students' learning progress and outcomes.

Furthermore, educators, closely involved with their students, are uniquely positioned to identify their specific needs and learning styles. With educators' observations about students' learning progress and challenges, LLM-based AIG can be adapted to create assessment items that are more effective in diagnosing students' learning gaps, misconceptions, and areas of strength. This can lead to more effective diagnostic items, providing valuable feedback for student learning and fostering targeted instructional strategies (Drori et al., 2022; Rodriguez-Torrealba et al., 2022). An interdisciplinary team can bridge the gap between technical development and educational application, increasing the accountability of using AI and leading to a more appropriate application of AIG in real-world educational and assessment contexts.

Another ethical consideration relates to the bias inherent in the content generated by LLMs. LLMs typically operate in a pre-training phase followed by either a fine-tuning or a prompting-tuning phase (Radford et al., 2018). In the pre-training phase, LLMs adopt an unsupervised learning strategy to learn the conditional probabilities of language tokens from vast text datasets. They can then undergo a subsequent fine-tuning phase to tune the parameters of the LLMs or receive task-relevant prompts without tuning the parameters to perform specific downstream tasks (P. Liu et al., 2023). Given that LLMs learn the conditional probabilities of texts from the training datasets, they often mirror the biases present in these datasets. Such inherent biases will propagate to downstream tasks, including their manifestation in the generated assessment items in AIG tasks. For instance, if an LLM is trained with datasets that predominantly feature texts associating certain demographic groups with specific characteristics, the model might generate texts that reflect and reinforce these social biases (Gallegos et al., 2024).

LLMs trained on data from low-resource languages or underrepresented groups may also generate items containing unfamiliar or inappropriate terms for these groups, raising fairness concerns (Bender et al., 2021). Such biases can manifest in item generation involving texts related to sensitive topics such as gender, race, religion, age, and nationality (J. Li et al., 2024). To address this ethical issue, several techniques can be implemented during multiple stages of training LLMs, including pre-processing, in-training, and post-processing (Gallegos et al., 2024). For instance, an effective method is the inclusion of diverse datasets in the preprocessing phase, enabling the model to learn from a wider array of contexts and reducing the likelihood of generating biased content (Buolamwini & Gebru, 2018). Moreover, model outputs can be evaluated and modified in the post-processing phase to mitigate bias, such as detecting and replacing harmful or inappropriate words. A more comprehensive introduction to the concepts of fairness and bias and the strategies for debiasing can be found in recent studies conducted by Gallegos et al. (2024) and J. Li et al. (2024).

2.1 Prompting for Item Generation

Inclusiveness stands out as one of the fundamental ethical principles in AI, encompassing the assurance of non-discrimination and the promotion of unbiased algorithms (Nguyen et al., 2023). Bias can be reinforced not only through the utilization of inherently partial data but also by incorporating instructions and cues containing biased language. Apart from the underlying algorithms or training data, the quality of prompts (i.e., specific cues and instructions provided to an AI model to generate desired content) employed for item generation plays a pivotal role in determining the efficacy of AI language models utilized in AIG (Bozkurt & Sharma, 2023). Prompt engineering is "the process of designing, crafting, and refining contextually appropriate inputs or questions to elicit specific types of responses or behaviors from an AI language model" (Bozkurt & Sharma, 2023). Heston and Khun (2023) also emphasize the capability of prompt engineering to adjust such aspects of the model's reply as "length, complexity, and style" (p.199). For instance, prompts can enhance item discrimination by eliciting brief, straightforward answers suitable for novice students while prompting more detailed responses tailored towards advanced learners (Heston & Khun, 2023). N. Zhang et al. (2021) also underscored the benefits of prompts, particularly their effectiveness in leveraging small datasets and their decreased reliance on specialized domain expertise.

Prompts characterized by vagueness or suggestive inclinations toward a specific response can yield biased outputs, and emotionally charged prompts can also impact the objectivity of responses and items generated (Heston & Khun, 2023). Such prompts may lead to generating content that perpetuates discrimination and prejudice. Moreover, using inaccurate or misleading prompts can produce false and deceptive information in AI models. To address this issue, Bozkurt and Sharma (2023) proposed a set of strategies for designing precise prompts necessary for obtaining the desired output. These recommendations entail setting clear objectives, employing appropriate language and tone, providing context, examples, and references, specifying the expected output format, and incorporating essential details anticipated in the response. Additionally, fine-tuning prompts, experimenting with different variations, and rigorously testing them to analyze results, along with adjusting prompts based on desired responses, can mitigate the generation of irrelevant or erroneous outputs and enhance overall performance (Bozkurt & Sharma, 2023).

Another common issue encountered in prompt engineering and item generation is explainability. The European European Commission (2019) stresses the significance of providing insights into how a specific AI algorithm operates and makes decisions. A lack of understanding of how prompts are constructed and applied can impede stakeholders' ability to effectively utilize assessments and cause potential misuse of AI capabilities (Nguyen et al., 2023). One of the solutions entails developing documentation that clearly outlines the purpose, provides guidance using prompt engineering techniques, offers examples of prompts, and showcases generated outcomes. These protocols can assist stakeholders in effectively applying prompts and understanding the mechanism behind item generation. By enhancing the clarity and transparency of prompts, their quality can be enhanced, thereby improving the validity and reliability of generated items (Lee et al., 2023).

Personalizing prompts with user-specific information can significantly enhance the relevance and quality of generated content, consequently contributing to a more positive learning experience. Strategies such as tailoring content based on previously collected user data or adjusting item tone, style, or complexity can lead to more personalized outputs. However, it is essential to ensure that these strategies are employed only with the consent of all parties whose data is utilized for prompt engineering. Privacy considerations that should be prioritized before implementation include obtaining consent, anonymizing data, ensuring data security, and maintaining transparency throughout the process (Nguyen et al., 2023; Vincent-Lancrin & Vlies, 2020). Prompt engineering holds considerable potential and can significantly enhance the quality of generated items. However, optimal design strategies should be developed based on ethical principles to mitigate these risks and ensure the integrity of the generated content (Heston & Khun, 2023).

2.2 Multimodal Stimulus Generation

Generating multimodal stimuli, such as images or audio, creates content that incorporates multiple modes of communication. Using graphics, audio, video, and interactive elements (e.g., biosensors) enables a more thorough evaluation of unscripted, complex tasks (Blikstein & Worsley, 2016). Multimodal assessments advance universal design principles in testing environments through multiple means of engagement, representation, and expression (Rao, 2015). These elements of universal design complement multimodal stimuli by ensuring that assessments are not only varied and engaging but also accessible to all learners, reducing barriers and enhancing the opportunity for every student to show their true ability. A study conducted by Smith et al. (2021) found that providing emergent bilingual students with multimodal assessments supported student identity expression.

By leveraging generative AI, it is possible to generate multimodal stimuli that can complement text-based questions is possible. For instance, in a language comprehension test, an AI could generate a short story as an audio clip alongside visual aids depicting key scenes or concepts. This approach not only aids in comprehension but also engages various cognitive skills, offering a more comprehensive assessment of the learner's abilities compared to traditional multiple-choice test items (Almond et al., 2010). Allowing students to interact with content in more meaningful ways facilitates a deeper measurement of their critical thinking, analysis, and interpretation skills (Sharma & Giannakos, 2020). This method can evaluate a broad array of skills and accommodate different learning styles. This ensures a more inclusive and effective assessment process (Sankey et al., 2010).

The integration of multimodal stimuli in assessments through generative AI raises several ethical considerations that are critical to ensuring the fairness and integrity of the testing process. One primary concern is the potential for bias in the AI algorithms used to generate these stimuli (Alwahaby et al., 2022). It is essential to ensure that these algorithms do not inadvertently favor certain groups of students over others based on cultural, socio-economic, or linguistic backgrounds. To identify and mitigate biases, it is necessary to conduct rigorous testing and maintain continuous oversight of AI systems. As previously stated, using a human-in-the-loop framework is crucial for multiple stages of assessment design, such as item quality review and sensitivity review (Hao et al., 2024).

Another ethical consideration is the privacy and security of the data used in creating multimodal stimuli (Alwahaby et al., 2022). As these assessments often incorporate personalized elements to cater to different learning styles, they might also collect sensitive information about students' preferences and abilities. Ensuring the confidentiality and secure handling of this data is paramount to maintaining trust and protecting students' rights. Also, the accessibility of multimodal assessments must be scrutinized to prevent the creation of new barriers to learning. While these assessments aim to be inclusive, there is a risk that the technology required to engage with them might not be equally available to all students, especially in lower-resource environments (Reiss, 2021). Institutions must provide adequate support and resources to ensure that all students have equal access to these innovative testing formats. Overall, while multimodal assessments offer significant advancements in measuring educational achievement, these technologies must be implemented thoughtfully and ethically to truly enhance the learning and assessment landscape.

3 Automated Scoring

Automated scoring of constructed-response items is one of the most successful early explorations of AI in assessment. Page (1966) and colleagues developed the Project Essay Grader (PEG) system, the first automated scoring system for essays. Automated scoring of short-answer items can be tracked back to the work by Burstein et al. (1998). AI scoring has also been used to score digitized spoken responses in language assessments (Higgins et al., 2011). NLP plays a critical role in processing text or speech data so that different machine learning models can be utilized to develop a scoring engine (Flor & Hao, 2021). The methods for automated scoring can be classified into two broad categories: the hand-engineered feature-based approach and the deep learning model-based approach (Haller et al., 2022; Lottridge et al., 2023). A hybrid of both has proven successful in recent years (Whitmer et al., 2023).

Feature-based models primarily utilize handcrafted features ranging from basic elements like word count to more complex aspects such as argumentation structure and coherence. These features include linguistic features such as syntactic, semantic, lexical, and readability, which are used as input to develop automated scoring models (Ke & Ng, 2019; Uto et al., 2020). The major modeling approaches (Ramesh & Sanampudi, 2022) include regression-based, classification models, and neural networks. In the 2021 National Assessment of Educational Progress (NAEP) Automated Scoring Challenge (Whitmer et al., 2023), the hybrid approach integrating handcrafted features and embeddings from BERT (Devlin et al., 2019) or extended BERT models performed slightly better compared with the pure ensemble models from LLMs models (Lottridge et al., 2023; Ormerod, 2022b, 2022a). Indeed, recent developments in AI scoring methods show a trend towards integrating LLMs, though use in operational testing is still not widespread, and concerns about validity and fairness persist.

3.1 Standards for Ethical AI Use in Automated Scoring

The Standards for Educational and Psychological Testing provide some guidance on AI scoring (AERA, APA, NCME, 2014). Standard 4.19 calls for well-documented support of the engine scores per score level, with theoretical and empirical bases that can link back to the construct. It also states: “The automated scoring algorithm should have empirical research support, such as agreement rates with human scorers, prior to operational use, as well as evidence that the scoring algorithms do not introduce systematic bias against some subgroups” (AERA, APA, NCME, 2014). This suggests the need for an evaluation of the automated scoring model as part of the evidence to support the use and interpretation of the scores. The Guidelines for Technology-Based Assessment of the International Test Commission (ITC) and the Association of Test Publishers (ATP) (ITC & ATP, 2022) offers a whole chapter on scoring, with an emphasis on constructed-response scoring, offering more detailed advice on AI scoring. In addition, two testing companies have published documents addressing ethical approaches to AI. The Best Practices for Constructed-Response Scoring published by ETS (McCaffrey et al., 2022) is focused solely on curating validity evidence for AI scoring and offers very detailed guidance for practitioners (including case studies). Duolingo’s Responsible AI Standards (Burstein et al., 2023) has a broad focus on the use of AI in assessment and offers high-level principles.

Given the complex nature of developing AI scoring systems, the extent to which the test users are provided with a layman’s-level description of the AI scoring engines and scores is vital. Guidance in the Standards (AERA, APA, NCME, 2014) predates the introduction of LLMs in educational testing. At that time, the concern with lack of transparency was related to NLP feature definitions and how they are combined to produce a score. Using LLMs creates a true “black box” with thousands or even millions of parameters to interpret. There are several approaches available to improve the interpretability and explainability of the AI scores (see Boulanger & Kumar, 2024; Molnar, 2020; Riordan, Bichler, Bradford, King Chen, et al., 2020; Riordan, Bichler, Bradford, & Linn, 2020; M. Zhang et al., 2024).

3.2 Bias in Automated Scoring

Historically, test bias has been evaluated by comparing scores from different demographic groups or subgroup populations. It may be intuitive to examine mean differences in group means or the overall impact; however, differential item functioning (DIF) analyses became the preferred method for detecting bias because it compares subgroups' item-level performance after conditioning on ability (Holland & Wainer, 1993; Shermis, 2024). Impact provides a measure of the overall total or scale score differences that are attributable to both actual differences in groups and differences brought about by the use of items with DIF (Angoff, 2012). AI scoring introduces a different source of potential unfairness into the testing environment that did not exist when DIF analyses and methods were developed.

To ensure that we can measure and address bias in educational testing, we must define the different types of bias and fairness (Johnson & McCaffrey, 2023). The literature on bias in machine learning (Mehrabi et al., 2021; Suresh & Guttag, 2021a, 2021b) focuses on seven sources of bias that are particularly relevant in the AI scoring context: historical bias, representation bias, measurement bias, aggregation bias, learning/algorithmic bias, evaluation bias, and deployment bias, and give example sources (Johnson & McCaffrey, 2023). For example, the features being used as part of the model may contribute additional biases if they carry construct irrelevant variance for some subgroups (Johnson & McCaffrey, 2023). Therefore, it is important to thoughtfully choose a model and features that best represent the task for the construct at hand and minimize construct irrelevance for all groups. Importantly, these various sources of bias could affect fairness in different ways. Recent literature (Johnson & McCaffrey, 2023) provides a nuanced conceptualization of fairness: AI scores can have independence (independent of group membership), separation fairness (conditionally independent of group membership, given the true score), sufficiency fairness (true score is conditionally independent of the group membership given the AI score), and/or conditional unbiasedness (i.e., test takers are not negatively affected by the use of machine scores or human ratings). Johnson and McCaffrey (2023) later discussed assessment-level fairness, which examines bias at the level of the reported score.

To minimize unfairness and develop a valid and accurate model, one of the largest considerations is to use a sufficiently large representative sample that includes all of the different demographic attributes of the larger population. Depending on the kind of assessment being performed and the engine used, significantly different sample sizes may be needed. Having more essays in the training data will broaden the set of response types and styles that may appear when the model is used for prediction. When the model contains many features, it will be essential to have larger samples to fully reflect the subgroups' feature distributions. In previous research, as few as 50 features have been found to create sufficiently accurate models (Burstein et al., 1998). However, obtaining sufficiently large sample sizes can be challenging in more contemporary models as LLMs generate thousands or millions of different features representing the test taker response. Features in the underlying language model should also be trained with a wide range of responses from all possible subgroups. See ETS (2021) for more detailed guidance on best practices for minimizing threats to validity and fairness.

3.3 Detection of Fairness Issues

Various approaches exist for detecting unfairness or checking to ensure adequate fairness in the context of automated scoring. D. M. Williamson et al. (2012) proposed a series of analyses to be conducted by subgroup. In practice, typically, only an analysis of standardized mean difference (SMD) comparing the human and machine scores for each group is performed (or reported). Other analyses may include a comparison of human-human Quadratic Weighted Kappa (QWK) to human-machine QWK by subgroup (Buzick et al., 2016) and DIF analyses (Shermis, 2024; Shermis et al., 2017; Vo et al., 2023). Differential feature functioning (Penfield, 2016; M. Zhang et al., 2017) analyses detect whether there are differences in engine feature performance across subgroups, conditioning on overall item score. Many studies report using a mixture of detection methods (He et al., 2022; Justice, 2022; Lottridge & Young, 2022). Johnson and McCaffrey (2023) provided methods to detect different types of fairness, including sufficient and separation fairness. In AI scoring systems based on generative AI with no human ratings, saliency methods may be crucial to understanding what aspects of a response lead to higher or lower scores. Recent work explores the use of saliency methods to detect and understand subgroup differences in short-response scoring (M. Zhang et al., 2024). Qualitative analyses performed by subject matter experts may also be required to provide evidence for validity in this context.

3.4 Potential Correction Approaches to Minimize Bias

To account for bias in feature selection, automated scoring engines can be developed with different sets of features or features weighted by subgroup. For example, in testing programs that use a contributory scoring approach, combining human ratings and machine scores (Breyer et al., 2017) for reporting, penalized best linear predictor models (PBLP) models may be used to minimize subgroup mean score differences between human and machine scores. Error-in-variables regression modeling (Johnson & McCaffrey, 2023) could identify features presenting issues for specific subgroups, and those features could be removed from the model. Another two modeling approaches were proposed to remove differences across groups: constrained optimization and direct penalization (Choi & Johnson, 2024; Johnson & McCaffrey, 2023). X. Liu and Fauss (2024) proposed a Bayesian non-parametric model for flexible automated scoring, which captures the potential nonlinear relationship between features and ratings for different subgroups.

Small sample sizes may jeopardize the representation of the characteristics of minority groups in developing automated scoring engines. Some data augmentation methods, such as oversampling (Chawla et al., 2002), can be applied to re-balance the small subgroups class in the training data. Synthetic text, audio, and video data can be generated by mimicking the characteristics of minority groups. This may enhance the representation of training data, ultimately minimizing potential sources of bias. However, measures should be in place to ensure that the response data generated are fair and unbiased.

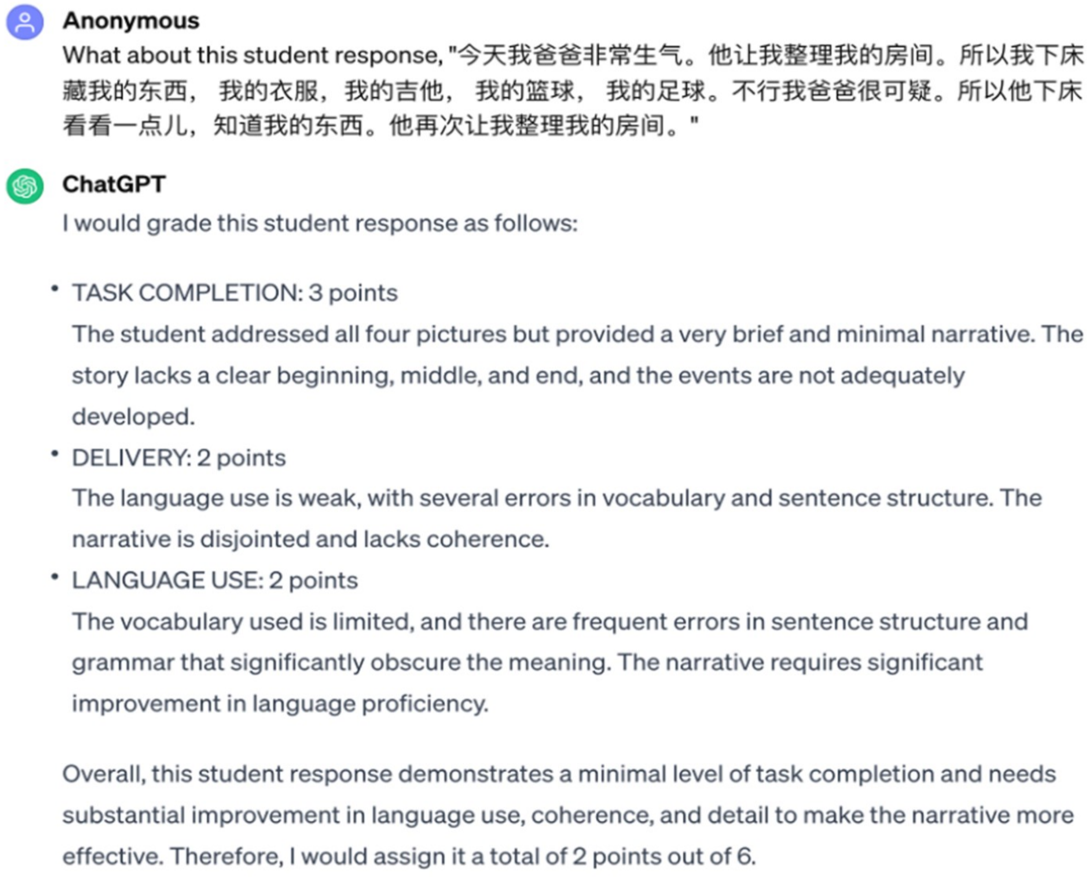
3.5 Illustrative Example: AP Chinese Language and Culture

To demonstrate the differences and possible ethical concerns between generative AI and traditional human scoring, we present below the results from a study comparing the scores of a human AP Chinese rater and two versions of ChatGPT (3.5 and 4.0). The essay prompts (scored on a 0 to 6 scale) were sourced from the 2021 AP Chinese Language and Culture Exam. Thirty third-year college students currently enrolled in a third-year Mandarin class participated. Each student was given 15 minutes to complete a story narration writing prompt based on four pictures and asked to imagine writing the story to a friend. In total, 30 essay samples were collected and graded by the human AP Chinese rater, as well as ChatGPT 3.5 and 4.0. Figures 1 shows a selected student writing sample below:

今天我爸爸非常生气。他让我整理我的房间。所以我下床藏我的东西，我的衣服，我的吉他，我的篮球，我的足球。不行我爸爸很可疑。所以他下床看看一点儿，知道我的东西。他再次让我整理我的房间。

Figure 1: A writing sample in Mandarin.

Before scoring, ChatGPT versions 3.5 and 4.0 were trained by a human rater on using the College Board's scoring rubric, with 10 example samples scored (0-6) by the human rater as references. Furthermore, all human and AI raters provided their rationale for each score based on the rubric. Figures 2 and 3 display the rationales provided by ChatGPT 3.5 and 4.0 for the selected student writing sample. Additionally, the human rater assigned a score of 3 to the same writing sample and cited the following rationale based on these criteria: "1) Task Completion: The narrative aligns with the stimulus but contains logical inconsistencies and disconnected sentences; 2) Delivery: The language usage is inconsistent and contains many errors; 3) Language Use: Limited vocabulary and idioms, with frequent errors that sometimes obscure meaning."



Anonymous

What about this student response, "今天我爸爸非常生气。他让我整理我的房间。所以我下床藏我的东西， 我的衣服，我的吉他， 我的篮球， 我的足球。不行我爸爸很可疑。所以他下床看看一点儿，知道我的东西。他再次让我整理我的房间。"

ChatGPT

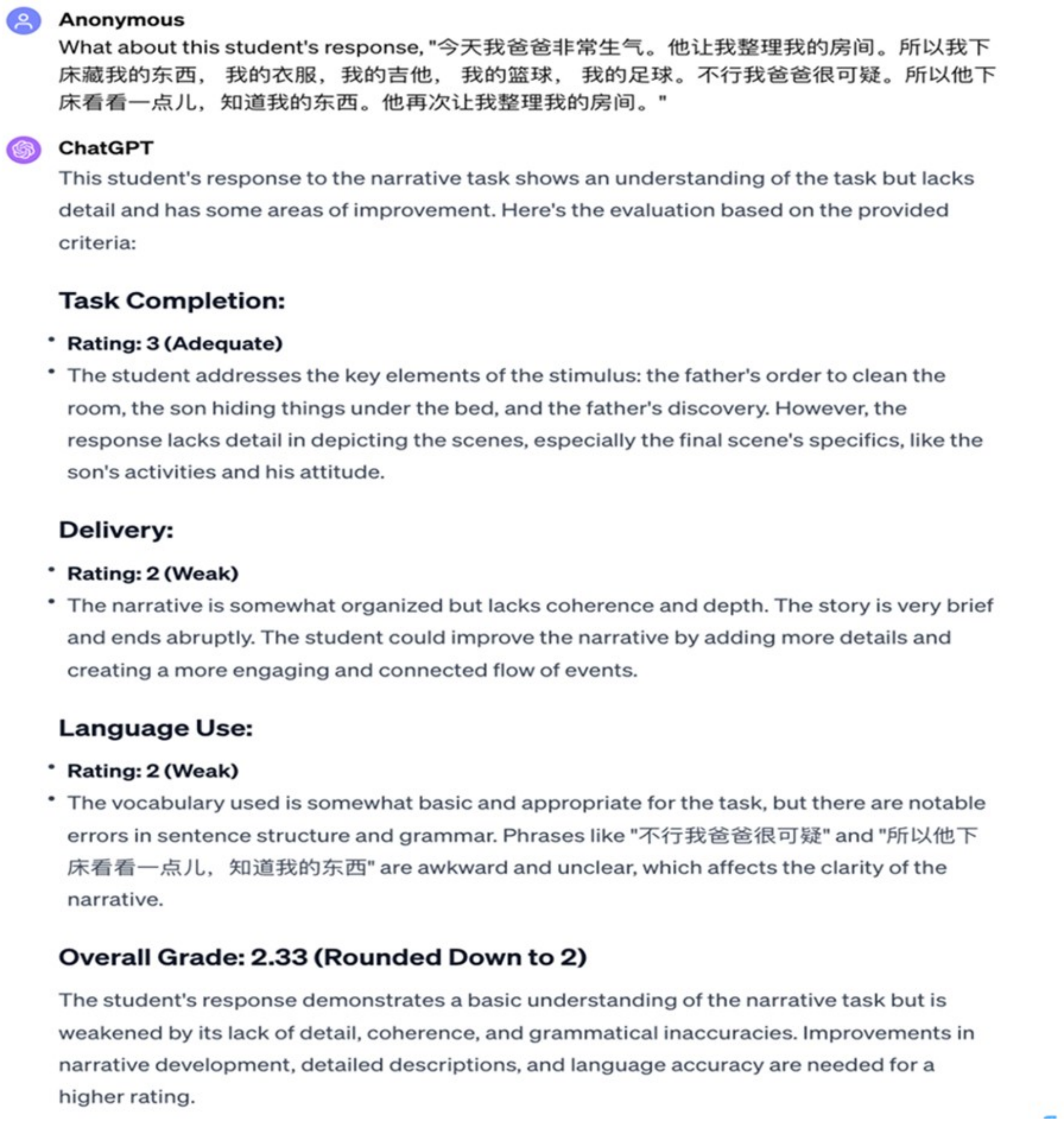
I would grade this student response as follows:

- TASK COMPLETION: 3 points
  The student addressed all four pictures but provided a very brief and minimal narrative. The story lacks a clear beginning, middle, and end, and the events are not adequately developed.
- DELIVERY: 2 points
  The language use is weak, with several errors in vocabulary and sentence structure. The narrative is disjointed and lacks coherence.
- LANGUAGE USE: 2 points
  The vocabulary used is limited, and there are frequent errors in sentence structure and grammar that significantly obscure the meaning. The narrative requires significant improvement in language proficiency.

Overall, this student response demonstrates a minimal level of task completion and needs substantial improvement in language use, coherence, and detail to make the narrative more effective. Therefore, I would assign it a total of 2 points out of 6.

Figure 2: Rationales provided by ChatGPT 3.5.

After reviewing all the rationales, we found that the scoring rationales provided by the human rater and ChatGPT were internally consistent. The human rater emphasized a holistic assessment, prioritizing task completion, followed by delivery and language use. In contrast, the ChatGPT models evaluated each criterion independently and averaged the scores. Despite their training to follow a similar prioritization as human raters, ChatGPT, in this instance, adhered to a methodology of averaging scores across domains. This methodological difference could potentially lead to a different distribution of scores, raising considerations about the validity evidence for the AI-based scores due to the distinct weighting given to each criterion.



Anonymous

What about this student's response, "今天我爸爸非常生气。他让我整理我的房间。所以我下床藏我的东西， 我的衣服，我的吉他， 我的篮球， 我的足球。不行我爸爸很可疑。所以他下床看看一点儿，知道我的东西。他再次让我整理我的房间。"

ChatGPT

This student's response to the narrative task shows an understanding of the task but lacks detail and has some areas of improvement. Here's the evaluation based on the provided criteria:

Task Completion:

- **Rating: 3 (Adequate)**
- The student addresses the key elements of the stimulus: the father's order to clean the room, the son hiding things under the bed, and the father's discovery. However, the response lacks detail in depicting the scenes, especially the final scene's specifics, like the son's activities and his attitude.

Delivery:

- **Rating: 2 (Weak)**
- The narrative is somewhat organized but lacks coherence and depth. The story is very brief and ends abruptly. The student could improve the narrative by adding more details and creating a more engaging and connected flow of events.

Language Use:

- **Rating: 2 (Weak)**
- The vocabulary used is somewhat basic and appropriate for the task, but there are notable errors in sentence structure and grammar. Phrases like "不行我爸爸很可疑" and "所以他下床看看一点儿，知道我的东西" are awkward and unclear, which affects the clarity of the narrative.

Overall Grade: 2.33 (Rounded Down to 2)

The student's response demonstrates a basic understanding of the narrative task but is weakened by its lack of detail, coherence, and grammatical inaccuracies. Improvements in narrative development, detailed descriptions, and language accuracy are needed for a higher rating.

Figure 3: Rationales provided by ChatGPT 4.0.

3.6 Illustrative Example: Comparison of Engines from NAEP Data Challenge

In 2021, the United States Department of Education's Institute of Education Sciences (IES) and the National Center for Education Statistics (NCES) hosted a data challenge to explore whether students' open-ended responses to the NAEP Reading assessment for fourth and eighth grade could be accurately and fairly scored using AI models. This first data challenge found that the top three challengers could, on average, accurately score all test questions with minimal degradation from the human-human QWK (human-human QWK – human-AI QWK < 0.05). The most accurate team overall used proprietary software called PEG, which uses an ensemble model that combines a series of classifier and regression machine models. The next two placing teams used different versions of BERT. However,

Item Name (Grade)	N Teams with Low QWK*	Mean QWK	SD QWK	Mean ELL SMD	SD ELL SMD	Max ELL SMD	Mean IEP SMD	SD IEP SMD	Max IEP SMD
Most Imp. Char. (8th)	4	0.78	0.07	0.18	0.04	0.21	0.04	0.03	0.10
Imp. Of Fast Deliv. (8th)	6	0.78	0.04	0.16	0.09	0.25	0.18	0.07	0.28
Text Box Helps (4th)	5	0.79	0.06	0.19	0.08	0.26	0.03	0.03	0.08
Thoreau Quotation (8th)	0	0.68	0.03	0.12	0.11	0.28	0.09	0.05	0.14
Innkeeper Changes (4th)	0	0.87	0.06	0.09	0.06	0.16	0.04	0.02	0.07
Describe Merchant (3th)	0	0.90	0.09	0.08	0.07	0.20	0.02	0.02	0.05
Reader Interested (8th)	1	0.88	0.13	0.09	0.09	0.24	0.05	0.05	0.11
I'm Ruined (8th)	5	0.79	0.10	0.04	0.03	0.09	0.11	0.05	0.18

Table 1: The teams with lower QWK had $\text{QWK}_{HH} - \text{QWK}_{HL} > 0.05$.

when the individual subgroup analyses were performed, there was some bias in predicted scores, mostly for English language learners (ELLs) and students with disabilities.

Table 1 shows the mean and standard deviation of the human-AI QWK and SMD in these groups across teams. The mean SMD for ELLs was 0.18 for the "Most Imp. Char." Task. The mean SMD for students with an individualized education plan (EIP) was 0.17 for the "Imp. Of Fast Deliv." Task. This level of differences in human and AI scores is typically considered large enough to be concerned, and further investigation is warranted (as per D. M. Williamson et al., 2012). Interestingly, the overall QWK for these two tasks was about 0.78 on average, indicating acceptable agreement. However, the tasks that showed higher human and AI agreement had very low mean SMDs. In general, when the overall model evaluation shows high levels of agreement and prediction accuracy (McCaffrey et al., 2024), there tend to be no subgroup differences (McCaffrey et al., 2022). This was found in the next challenge in 2023 when NCES/IES hosted a challenge to explore the accuracy and fairness of scoring open-ended responses to the NAEP Math assessment for fourth and eighth grade. The top three teams were as accurate as the human raters on average across all ten items that were scored—they used completely different methods but yielded a high average QWK (> 0.93). There was no demonstrated bias in any of the major demographic groups.

Automated scoring is not at all a new capability in educational assessment, with the first systems created in the 1960s. These earlier systems were not based on AI, per se, but on supervised models built by NLP scientists using feature sets that were somewhat explainable. With the recent explosion of AI in the mainstream and the availability of open-source advanced LLMs, there has been a resurgence of interest in automated scoring. AI is now more broadly applied to advance and potentially improve automated scoring of constructed-response items. Along with this increase comes the potential responsibility to investigate how the new technology functions, particularly validity and fairness issues. This section aimed to acquaint readers with some of the topics to consider and to encourage more detailed exploration into the current research being done in our field to address validity and fairness in AI scoring of constructed-response items.

4 Score Reporting and Feedback

In educational measurement, feedback serves as a crucial component that transforms examinees' performance into actionable insights. Its purpose is to inform individuals about their current ability levels and provide them with guidelines to enhance their knowledge for improved future learning outcomes (Boud & Molloy, 2013). Feedback contributes to students' understanding of their performance by operating on multiple levels as follows (Hattie & Timperley, 2007):

1. The *task* level involves guiding students on what specific actions are needed to arrive at the correct answer.
2. The *process* level focuses on informing students about the thought processes required to approach tasks correctly.
3. Self-regulated learning (SRL) level pertains to the ability of students to monitor and guide themselves toward achieving their learning objectives.
4. The *self* level includes reflective feedback of the learners themselves as individuals.

Feedback in its various forms addresses different levels of change. For example, offering scores and specific instructions related to tasks can stimulate modifications at the task level. However, to foster SRL, feedback must be contextualized to the students' circumstances, thereby promoting introspection on their task approach (Carless, 2019). An instance of feedback at the SRL level could be, "The strategy you used to approach the task did not do well. What do you think went wrong? What else can you do to reach a different outcome?"

4.1 AI's Role in Feedback Enhancement

AI-powered technologies can assist educators in formulating effective feedback through information gained from learning analytics (LA) and NLP (Wongvorachan et al., 2022). LA can provide insights into students' performance and learning by analyzing data about students. Simultaneously, NLP can automatically translate numerical components into actionable verbal feedback. This cutting-edge capability could significantly reduce instructors' workload in the case of large class sizes and the influx of information. Especially in this era where online- and hybrid learning is prevalent, AI can leverage the wealth of information generated from online learning platforms such as Moodle (Jin, 2012). This capability not only informs students of how they did but also informs instructors about patterns of students' learning (e.g., distribution of quiz scores) to inform their instruction planning as well.

The mentioned innovation is driven primarily by students' data, such as students' engagement with course materials through a learning management system and their assessment scores (Bulut, Gorgun, Yildirim-Erbasli, et al., 2023; Yildirim-Erbasli, Gorgun, & Bulut, 2023). For this reason, instructors and researchers must ensure that all information is given with consent and that no misuse occurs (Stahl & Karger, 2016). Schools and academic institutions should implement data privacy measures, educating students about their data rights and maintaining transparency about how and why their data is being used (Lin & Kifer, 2015; Stahl & Karger, 2016). Potential misuse of students' data in the feedback process includes data misinterpretation, inconsistent data definition (i.e., comparing different kinds of data), poor result visualization (i.e., choice of graphs, texts, and colors) (Webber, 2019). For example, results about students' performance, including the overall score distribution of the class, should be given with context. Inferring the entire semester's performance from midterm scores, meant for formative assessment, is inappropriate. Such data misuse may inadvertently discourage students from taking initiative in their learning, shifting the focus to test-taking proficiency rather than genuine knowledge acquisition (Jones & Egley, 2007).

Introducing AI into the feedback process presents an opportunity to enhance and streamline educational practices. AI can automate the feedback process, provide personalized insights, identify learning gaps, and adapt to the learning pace of each student. This can lead to more efficient learning experiences and improved educational outcomes. However, while harnessing the power of AI, it is crucial that ethical considerations are accounted for. The use of AI in education involves handling sensitive student data, which may cause harm to students if misused. Educators and researchers must ensure that all data-driven insights are derived and utilized responsibly. This means being transparent about data collection and usage policies, implementing robust data privacy and security measures, and regularly reviewing and updating these measures in line with evolving ethical standards and regulations (Gimpel et al., 2018).

4.2 Illustrative Examples: Automated Writing Evaluation

Automated writing evaluation (AWE) has evolved significantly since its initial discussion in the scholarly literature by Warschauer and Ware in 2006, who set a foundational research agenda for its development. AWE now broadly encompasses tools that provide both quantitative scores and qualitative feedback within classroom-based formative assessments (Hockly, 2019; Huawei & Aryadoust, 2023), although some AWE tools solely provide qualitative feedback. Automated scores and feedback are generated using algorithms that rely on NLP and AI—primarily supervised machine learning but increasingly LLMs—to deliver automated feedback aimed at enhancing student writing (Correnti et al., 2024; Cotos, 2023; Deeva et al., 2021; Fu et al., 2024; Shi et al., 2022). Linked initially with automated essay scoring (AES) and automated essay evaluation (AEE) for shorter constructed-response items, AWE has expanded to support a broader range of writing activities. It offers utilities that assist with the entire writing process, including planning tools like concept mapping and outlining, alongside feedback mechanisms that focus on grammar, organization, and development (Burstein et al., 2018; Huawei & Aryadoust, 2023). AWE is intended to allow educators and learners to customize feedback, aligning it with individual writing goals and pedagogical objectives.

AWE has demonstrated notable benefits in improving writing quality across various contexts, with studies reporting effect sizes ranging from 0.38 to 0.98 in tertiary and second language (L2) settings (Fleckenstein et al., 2023; R. Li, 2023; Ngo et al., 2024; Nunes et al., 2022). Despite these positive outcomes, the effectiveness of AWE varies significantly by educational level, duration of intervention, and the nature of feedback provided. For instance, AWE tends to be less effective for younger students, such as middle schoolers, particularly when feedback is generic or lacks integration with comprehensive writing curricula (Ware, 2014; Nunes et al., 2022). Moreover, while some studies indicate that AWE can boost motivation and self-efficacy in writing (Grimes & Warschauer, 2010; Moore & MacArthur, 2016; Warschauer & Grimes, 2008; Wilson & Roscoe, 2020), others note mixed effects on students' writing-related beliefs and affect, with variability often tied to the method of AWE implementation and students' prior skills levels (C.-F. E. Chen & Cheng, 2008; Wilson, Zhang, et al., 2024). For example, high-achieving students may experience demotivation when using AWE because the system does not provide as positive an evaluation of their work as their teacher does (Wilson et al., 2021).

Given that AWE has differential effects based on how it is implemented, educators, developers, and researchers should carefully consider how AWE is most ethically, equitably, and effectively deployed. These considerations encompass the representativeness of data, the transparency and explainability of algorithms, the assurance of equitable access and benefits across diverse student demographics, and maintaining human oversight. Addressing these factors is essential to ensure that AWE tools support educational objectives ethically, fostering an inclusive and fair learning and assessment environment.

4.2.1 Representativeness of the Training Data

Ensuring data representativeness in AWE systems is a critical ethical consideration that impacts the fairness and effectiveness of these tools. Developers must assemble training data that reflect the diversity of the intended user base, not only in demographic terms but also across the spectrum of achievement levels expected among users (see Raczynski & Cohen, 2018). This involves including a wide range of essays, scores, and feedback that capture both typical and atypical performance outputs—often underrepresented at the extremes of the achievement distribution. In addition, it is crucial to ensure that human ratings and feedback are scrutinized for bias, including rater biases (e.g., Wind et al., 2018), as well as human biases, stereotypes, and language ideologies (Goldshtein et al., 2024). Such comprehensive data collection helps train AWE systems that are unbiased and equitable, capable of providing relevant and accurate feedback to all users, regardless of their background or initial skill level. At present, it is not common practice to divulge details about the training data underlying AWE models, but this may be a practice the field pushes to adopt.

4.2.2 Transparency and Explainability

Transparency and explainability in AWE systems are crucial for building trust and enhancing usability among both students and educators (Goldshtein et al., 2024). These elements ensure that users understand how the AI generates feedback and the rationale behind the specific suggestions offered. As Myers and Wilson (2023) discuss, transparent AWE systems provide detailed explanations of their operational mechanisms, allowing users to see the connection between their input (the student's text) and the output (the feedback provided). This clarity helps demystify the AI processes, which is essential for users to confidently rely on and effectively utilize the feedback. Indeed, a central premise of formative assessment is clarifying and making explicit for learners the criteria for success (Black & Wiliam, 2009). In writing, knowledge of evaluation criteria serves as the basis for revising in the absence of automated feedback. Moreover, when AWE systems clearly articulate the basis of their feedback, they are more likely to be integrated successfully into educational practices, as both students and educators can better align the automated feedback with instructional objectives and learning outcomes. Finally, ensuring that AWE systems are transparent also aids in accountability, making it easier to facilitate third-party research and evaluation and to identify and correct potential biases or errors in the AI's assessments. This level of transparency not only improves the educational tool's efficacy but also fosters a deeper trust in technology-enhanced learning environments (Myers & Wilson, 2023).

4.2.3 Equity and Access

Ensuring equity in access and benefits from AWE systems is a pivotal ethical consideration, particularly as these technologies become more integrated into educational environments. Equity challenges in AWE usage stem from the need to ensure that all demographic groups, especially vulnerable or under-served populations such as ELLs, receive comparable benefits from these technologies. This involves more than just making AWE tools available; it requires that these tools are effective across diverse linguistic, cultural, and socioeconomic backgrounds. For instance, the challenge of equitable access and benefit in AWE systems includes ensuring that the feedback provided is linguistically and culturally responsive. This means that the systems must be capable of recognizing and adequately responding to the varied language use among students from different backgrounds.

Additionally, AWE systems should be tested and proven effective in diverse educational settings to avoid perpetuating existing educational disparities. A recent study by Wilson, Palermo, and Wibowo (2024) illustrates an approach to conducting such testing. They examined whether elementary-aged ELLs and non-ELLs equally solicited AWE feedback, improved the quality of their first drafts, and productively revised their writing. Findings were promising: language status was unrelated to the degree to which elementary-grade students accessed and benefited from AWE. However, this type of nuanced research is not yet commonplace. Such testing should become the norm, supplementing the rigorous psychometric testing of the underlying scoring and feedback algorithms.

4.2.4 Human Oversight

AWE is not designed to, nor should it, replace the teacher. As articulated in human-centered AI principles (Shneiderman, 2022), the deployment of AWE should enhance the teacher's (and student's) capabilities. For example, educators using AWE have been shown to focus more on complex writing skills by offloading the evaluation of basic and generic skills to the AWE system. Research supports this approach, indicating that AWE enables teachers to provide richer, more focused feedback on advanced writing elements (Wilson & Czik, 2016). However, unlike insights from teachers or peers, AWE feedback lacks contextual awareness of a student's developmental progress or curriculum timeline. It analyzes students' texts against the corpus of texts in its training data and the algorithms built thereupon. Thus, AWE may provide feedback on skills students have not yet encountered or are not expected to master at their current educational stage. This situation underscores the importance of teachers' ongoing involvement in correcting any misalignment and ensuring the AWE's feedback supports rather than contradicts their pedagogical objectives. This dynamic was highlighted in focus groups with teachers who have integrated AWE into their classrooms (Wilson et al., 2021). Moreover, it is crucial for students to understand that AWE systems are tools intended to enhance their ability to communicate effectively with humans rather than serving as the ultimate judge of their writing proficiency. Educators must ensure that AWE is used as an aid in the broader context of developing competent communicators, emphasizing that the technology should complement, not dictate, the learning process.

It is clear that AWE's potential and its challenges are closely tied. The advancements in AWE technology offer opportunities to enrich educational practices by providing timely, consistent feedback and freeing educators to focus on higher-level teaching objectives. However, these benefits hinge on addressing critical ethical considerations, such as ensuring the representativeness of training data, maintaining transparency and explainability, promoting equity in access and benefits across all student demographics, and ensuring that teachers remain in the loop. As AWE systems begin to integrate powerful generative AI technologies—technologies that are more opaque than AI used in legacy AWE systems—these ethical considerations will become all the more salient.

5 Other Concerns on AI Use in Education

5.1 Aberrant Responses

In an assessment, aberrant response patterns may occur due to atypical test-taking behaviors such as cheating, careless responding, creative responding, and non-effortful responding (Bulut, Yildirim-Erbasli, & Gorgun, 2024; Gorgun & Bulut, 2021; Kim et al., 2016; Wan & Keller, 2023). Such behaviors can arise in both high-stakes and low-stakes assessment contexts (Y. Liu et al., 2020). For instance, in high-stakes assessments, some test-takers may be motivated to increase their scores by engaging in dishonest behavior (Ranger et al., 2023). Conversely, non-effortful responding tends to be more prevalent in low-stakes assessments, where students may be less motivated to sustain sufficient effort throughout the test to demonstrate their true ability (Lindner et al., 2019; Wise & Kong, 2005).

Aberrant responses pose a significant concern for educational measurement because of their impact on data quality and the validity of inferences or predictions made using assessment results (Gorgun & Bulut, 2021; Kim et al., 2016). Thus, much research has been devoted to detecting and handling aberrant responses using different techniques. Notably, in recent years, machine learning approaches have been developed to complement psychometric approaches (e.g., person-fit indices, response time models) in this endeavor, making it possible to investigate aberrant response behavior using multiple sources of data (Kim et al., 2016; Mueller et al., 2016).

Regarding the detection of non-effortful or disengaged responses, this is usually operationalized by examining response time data collected from digital assessments, seeking to identify instances where test-takers spend unrealistic amounts of time (either too long or too short) on items or tasks (Gorgun & Bulut, 2021; Lindner et al., 2019; Y. Liu et al., 2020; Yildirim-Erbasli & Bulut, 2022). Typically, the data would not include the ground truth about aberrant responses, so the problem is approached from an outlier or anomaly detection perspective. Some approaches include threshold-based methods (e.g., Soland et al., 2021) and mixture modeling (e.g., Y. Liu et al., 2020; Wang et al., 2018). Drawing upon the capabilities of machine learning, researchers have also explored the utility of supervised (e.g., Yildirim-Erbasli & Bulut, 2022) and unsupervised learning algorithms, often incorporating additional features other than response time (e.g., Gorgun & Bulut, 2022).

Another area of particular interest is cheating detection. Over the last decade, researchers have applied and tested various machine learning and deep learning algorithms to detect cheating behavior, including supervised and unsupervised learning approaches (Cizek & Wollack, 2016; Jiao et al., 2023). In terms of data input, item-level data (i.e., responses and scores) are most often used to detect aberrant response patterns. Kamalov et al. (2021) presented an algorithm using sequences of grades from students' continuous assessment results to identify cheating on the final exam. More recently, the availability of process data has spurred further advancements. Process data are collected throughout the course of a digital assessment, such as response time, frequency of item revisits, clickstream data, and even sensor data tracking eye movements or head positions (Alsabhan, 2023). Many researchers (e.g., Alsabhan, 2023; Meng & Ma, 2023; Ranger et al., 2023, Tang, 2023, T. Zhou & Jiao, 2023) have demonstrated how process data could be leveraged to improve the performance of machine learning algorithms.

In the guidelines on quality control in scoring, test analysis, and reporting of test scores set out by the International Test Commission (2014), it is recommended that aberrant response patterns should be routinely monitored to uphold test security and integrity. With the increasing availability of data from computer-based testing systems, AI no doubt offers a promising solution to enhancing the detection of such patterns. There are several ethical considerations to take into account. From a social perspective, we must be careful about how results from detection algorithms are used in order to avoid unintended consequences. This ties into the Responsibility principle put forth by Taiwo et al. (2023). As Kim et al. (2016) stated, "The use of data to identify statistically improbable test behaviors is defensible. However, generalizing the results to a testing individual or population to prove cheating is problematic." (p. 71). In this way, the goal of aberrant response detection should be related to the validity of the scores and not cheater identification (Kim et al., 2016). Suppose results indeed hold consequences for individuals (e.g., score invalidation). In that case, it is paramount that these decisions are explainable and defendable and clear communication channels are set up with test-takers (Mueller et al., 2016).

Aberrant response behavior is often of great interest to stakeholders in educational assessment, such as test developers, test sponsors, and educators. A vast body of research has been built around methods to improve its detection. While each detection method has its own merits, Mueller et al. (2016) recommended that it is time for researchers to think more collectively and draw links between different indicators. This means working towards developing a process that identifies aberrant responses using multiple indicators rather than relying on evidence from a single technique (Mueller et al., 2016). The development of AI algorithms supports this vision, as multiple sources of data could be incorporated into the same model. However, the challenge lies in its transparency and explainability (another ethical principle outlined by Taiwo et al., 2023). Fairness and potential algorithmic bias must also be considered, especially when background and demographic variables are included as part of the data input.

5.2 Predictive Utility of Assessment Results

Beyond providing information on students' performance, AI-powered innovations can utilize student data, including test results, to predict various educational outcomes. For instance, a student's scores from formative assessments and the time taken to complete tasks have been identified as strong predictors of their summative assessment scores (Bulut & Yildirim-Erbasli, 2022). Similarly, a student's Grade Point Average (GPA) in the ninth grade can predict their likelihood of high school dropout (Bulut, Wongvorachan, et al., 2024). These examples demonstrate that pedagogically grounded predictors, such as assessment-related data, can provide more actionable insights than non-pedagogical predictors like socioeconomic status or complex predictors, such as clickstream data. This application of AI in predictive tasks falls under the domain of Educational Data Mining (EDM), which involves extracting knowledge from educational databases (Wongvorachan et al., 2022). While LA and EDM both use educational database variables to inform students and instructors, EDM distinguishes itself by prioritizing the optimization of predictive models for accurate predictions (G. Chen et al., 2020).

EDM can leverage students' assessment results, among other variables, to predict potential future outcomes. This information can be used for early intervention by informing parents, teachers, and students. For instance, a student's first-generation status and American College Testing (ACT) scores, in conjunction with their GPA, can predict university retention (Trivedi, 2022). Furthermore, a student's career prospects can be predicted using their GPA and performance in mock interviews, assessing factors like self-confidence, presentation ability, and communication skills (Casuat & Festijo, 2019). These insights can guide students and parents in future preparation. Instructors and administrators can use these results to initiate student support programs, such as remedial classes, writing support centers, or career counseling services.

While EDM is helpful in this regard, it is crucial to consider the explainability and actionability of its predictions. Without these, predictions may not be useful, as we can only anticipate outcomes without the ability to act upon them. To address this, the application of Explainable AI (XAI) is essential. XAI, a branch of AI that focuses on making the output of complex predictive models understandable to humans, helps establish trust between the user (e.g., instructors) and the tool (i.e., the model) (Biecek & Burzykowski, 2021). Specifically, XAI can provide global-level explanations to identify influential predictors through variable importance methods and local-level explanations to elucidate the mechanism behind a case's prediction through methods like Local Interpretable Model-Agnostic Explanations (Biecek & Burzykowski, 2021). By making the results understandable, we enable human validation of the prediction results to confirm their alignment with reality before taking actionable measures (Bulut, Wongvorachan, et al., 2024).

5.3 AI-Powered Proctoring

Another useful application of AI in assessment regards online test administration. Online learning and examinations were on the rise even before the onset of the COVID-19 pandemic when Massive Open Online Courses, also known as MOOCs, and colleges were leveraging online tools to provide more flexible access to resources for students. The COVID-19 pandemic then forced many educational institutions across the globe to rapidly accelerate this investment in online learning and examinations (Moreno-Guerrero et al., 2020; Nigam et al., 2021). The benefits of conducting exams online as opposed to the traditional in-person format include the ease of exam scheduling and asynchronous exam administration. There is also no need to find a physical space to conduct an exam, allowing exams to be conducted at massive scales without worrying about student capacity (Arora, 2021). Online exams require remote proctoring services, however, which come with one significant drawback: the ratio of human proctors to test takers is typically higher in online formats because online exams provide more opportunities for academic misconduct, which leads to the need for more proctors to effectively invigilate the exams (Bilen & Matros, 2021). One solution to this problem comes in the form of AI-based proctoring systems, which can help alleviate the human proctoring burden.

Remote proctoring may rely on a variety of AI tools, including face detection, eye gaze detection, keystroke analysis, lockdown software, web traffic recording, and others (Dyer, 2024). Typically, AI-based proctoring starts with identity verification, where the system confirms the test taker's identity through biometrics such as facial recognition and voice recognition or even IP address verification to confirm the test taker's location. Using facial recognition, the test taker's image is taken immediately prior to starting an exam and compared to some verifiable identification, such as a school ID card, to confirm whether the person sitting for an exam is indeed the person who is registered for the exam. This verification procedure can occur periodically throughout an exam to confirm that the registered test taker is taking the exam at any given instance (Nigam et al., 2021). Beyond facial recognition, facial detection can be used to identify how many people are present in an image and flag instances where no faces or more than one face was detected (Motwani et al., 2021).

Similarly, gaze tracking can be used to identify where a test taker's attention is focused and if they direct their attention away from their screen to other resources such as external notes or a second screen (Singh et al., 2022). An AI proctoring system can employ object recognition alongside this functionality to identify what test takers are directing their attention to (if it is within the camera's field of view) and flag whether the test taker uses non-permitted objects (e.g., calculators). When it comes to audio, AI can also be used to identify any background noise, which can be analyzed to determine whether a test taker received unpermitted assistance in completing their exam either from another individual in the room outside of the camera's field of view or from someone on another call with the test taker (Nigam et al., 2021).

Despite its benefits, AI proctoring also comes with several issues. First, facial recognition and AI audio detection tools are not exempt from producing false positives and can, therefore, mistakenly assign such flags for academic dishonesty (Slusky, 2020; Nigam et al., 2021). So, while AI proctoring can minimize human proctoring efforts, human review of these flags is still necessary (Dyer, 2024) to prevent unfairly penalizing students. Additionally, false negatives are also a concern, and there is little peer-reviewed evidence on the efficacy of remote proctoring in detecting cheating (Dawson, 2024). In a controlled study on a particular proctoring service, where six out of 30 students were asked to cheat, Bergmans et al. (2021) found that the remote proctoring software did not flag any of the six students. In comparison, a human review of the video recordings led to catching one of the six students who were asked to cheat.

Second, not only can false positives impact the academic standing of test-takers, but they are also more likely to occur for particular demographic groups; as any tool dependent on AI algorithms, remote proctoring is not free from bias. For instance, in a study on automated proctoring software, Yoder-Himes et al. (2022) found evidence of race, skin tone, and gender bias in the facial detection algorithm, which was significantly more likely to flag women with darker skin tones for review than men or women with lighter skin tones.

Third, in addition to concerns around AI accuracy, reliability, and fairness, there are also key considerations regarding data privacy and security. Since AI proctoring typically involves some form of analyzing biometric data and often even video feed from the test-taker environment, it can also be perceived as an invasion of privacy (Coghlan et al., 2021), which could sometimes lead to legal action (Dyer, 2024). With respect to data security, companies that provide remote proctoring services and the institutions that use them must comply with legal guidelines (e.g., GDPR) and have systems in place that can prevent sharing this data with 3rd parties and mitigate the risks of potential cyber-attacks. However, as Coghlan et al. (2021) pointed out, and as in the case of any AI tool, these technical controls are necessary (Slusky, 2020) but not failproof.

Lastly, given that AI proctoring often relies on the test-taker equipment (e.g., the student's laptop webcam), this can lead to different technical issues individual students may face, depending on their equipment and its compatibility with the proctoring and test-taking platform. For this reason, despite the various advances in AI proctoring, online proctoring would also preferably be supplemented by a technical support team that could assist test-takers in the eventuality of technical difficulties during the test. Indeed, even detecting these technical difficulties is made easier with AI. For instance, facial recognition software could flag webcam issues, while AI audio detection could flag whether there is no audio feed for a test that might require the student to speak. Nevertheless, human review of these flags would still be required (Nigam et al., 2021), and ideally, a combination of AI and human live remote proctoring would be employed (Dyer, 2024).

Some of the risks listed above could be mitigated by employing a hybrid remote proctoring model that uses both AI detection and live human proctors. Alternatively, the needed human review of AI-assigned flags could happen asynchronously, with the recordings being reviewed to determine whether an academic integrity violation occurred. However, research conducted by one company that provides remote proctoring services found that videos flagged by AI systems due to irregular testing behavior underwent human review around 10% of the time (Jaschik, 2021). This finding led the company to no longer offer remote proctoring solely based on AI and move towards a hybrid remote proctoring model.

Irrespective of the type of AI-assisted proctoring employed, the institutions that rely on these services must establish clear guidelines on the use of such tools and procedures for following up on potential academic integrity violations. Alongside clear guidelines, best practices for the ethical implementation of remote proctoring also include effective communication and training of faculty and staff (Dyer, 2024). The goal is for everyone involved in any remote proctoring system to be knowledgeable and able to speak to the services carried out by the system. In this sense, creating an internal remote proctoring implementation team could help by training educators and staff who could then discuss these systems with students. Regular communication between faculty and the remote proctoring team can also help an institution adapt quickly and make changes to the implementation of a remote proctoring system based on ongoing feedback. Crucially, decisions about AI proctoring tools need to be transparently communicated to both educators and students, ensuring that students know what to expect and what will happen during their exams and preparing them for a successful test-taking experience. Lastly, students should also be provided with additional options in case they are unwilling or unable to participate in a remote exam, and this process for requesting an alternative arrangement should be communicated clearly.

5.4 Automation Bias

Automation (i.e., the transition from human labor to computerization and mechanization; Rahm, 2023) through AI-based tools plays a critical role in aiding both educators (e.g., grading and analyzing student data) and students (e.g., using automated feedback systems to guide their learning and progress) (B. Williamson et al., 2023). Designers and developers of AI-based assessment tools argue that these tools can revolutionize and better educational assessments and student experiences (e.g., Bulut & Yildirim-Erbasli, 2022; Yildirim-Erbasli & Bulut, 2023; Yildirim-Erbasli, Bulut, et al., 2023; Yildirim-Erbasli & Bulut, 2022). For example, the integration of AI-based tools can offer the promise of minimizing decision errors, such as grading where human error can occur D. M. Williamson et al. (2012); M. Zhang (2013). When functioning effectively, this automation can enhance the accuracy and efficiency of assessment processes, benefiting both educators and students. However, this adoption of automation introduces its own possible errors (Lyell & Coiera, 2017; Jones-Jang & Park, 2023). For instance, automated grading systems may inadvertently misinterpret students' written responses or fail to capture the nuance of complex concepts, leading to inaccuracies in assessment outcomes when they rely on primary factors like essay length (Andersen et al., 2021). Regardless of the accuracy of these tools, the overreliance on automation can lead to errors in decision-making. Automation bias denotes the phenomenon where decisions are influenced by an overreliance or excessive dependence on AI-based systems, even when these systems may be flawed or incorrect (Bond et al., 2019; Lyell & Coiera, 2017; Parasuraman & Riley, 1997). For example, a student unquestioningly accepts a high grade from an AI-based grading system without considering the validity of the feedback or their own understanding of the material.

One prominent issue about automation bias is blindly accepting AI-based tools' outcomes without critical examination (e.g., Khera et al., 2023; Kupfer et al., 2023). When educators place too much trust in AI-based tools, they may overlook errors or biases inherent in these systems. This lack of critical scrutiny can result in unjust outcomes for students, as their performance may be inaccurately or unfairly represented. Another concern is to stifle students' skills (B. Williamson et al., 2023). If educators become overly reliant on automated assessment tools, they may prioritize conformity to standardized metrics over fostering essential skills such as critical thinking, creativity, and problem-solving. This narrow focus on quantifiable outcomes could limit students' intellectual development and inhibit their ability to thrive in a rapidly changing world. Furthermore, automation bias can raise questions about accountability and transparency (B. Williamson et al., 2023). When decisions about student performance are delegated to automated systems, it can be challenging to hold stakeholders or institutions accountable for errors or biases in the assessment process.

In addition, overreliance on automated assessment systems may reduce human interaction and feedback, depriving students of valuable opportunities for personalized guidance and mentorship: concern about dehumanization (Fritts & Cabrera, 2021). Moreover, suppose students become accustomed to receiving automated grading and feedback without understanding the underlying reasoning. In that case, they may become less adept at critically evaluating their work and identifying improvement areas. Second, there might be the issue of unquestioning acceptance of feedback. Automation bias may lead students to accept feedback provided by AI-based tools without critically evaluating its accuracy or considering alternative perspectives. This uncritical acceptance could hinder students' development of essential self-assessment and self-regulation skills. Stakeholders of educational assessments should ensure that automation serves as a valuable tool for enhancing, rather than detracting from, the assessment experience for all students.

Ethical discussions around automation bias extend to considerations of education and training (Rahm, 2023). It is important to educate students and other stakeholders in educational assessments about the limitations of automated systems and empower them to critically evaluate and supplement automated outputs with human judgment when necessary. In addition, human-centered AI has been discussed as a means to mitigate the automation bias. Human-centered AI places humans at the center of the design process, focusing on creating, developing, and deploying AI-based tools that prioritize human values, needs, and experiences (Bond et al., 2019). Following human-centered AI, AI-based assessment tools should have the capability to steer users (e.g., students and educators) through the decision-making process, enabling individuals to make their own micro-decisions (see Cairns et al., 2016). AI decisions can be reserved until the conclusion, serving to mitigate automation bias and anchoring effects from the outset. The AI tools can also present multiple competing decision statements and explanations to enhance transparency and encourage the user to engage in reasoning and arrive at the final decision, referred to as a differential diagnosis (Bond et al., 2019).

5.5 Evidence-Based Practice in Educational Assessment

Evidence-based practice refers to the intentional, reliable, and judicious use of empirical evidence to inform real-life decisions. In the context of educational assessment, evidence-based practice serves as the cornerstone for ensuring the efficacy, fairness, and reliability of assessment applications. It involves a deliberate and meticulous approach to decision-making rooted in robust research findings and empirical evidence.

Empirical evidence holds immense significance, especially in the context of AI-powered assessment tools, for several reasons. First, it acts as a safeguard against different sources of bias and discrimination. By rigorously examining data gathered from diverse populations, researchers can identify and mitigate potential biases embedded within algorithms or AI-based assessment frameworks. This process is crucial for ensuring that AI assessments provide equitable opportunities for all individuals, regardless of their background or characteristics.

Second, empirical evidence allows for the continuous refinement and improvement of AI-powered assessment tools. Through ongoing research and data analysis, researchers and assessment experts can identify areas of strength and weakness within AI algorithms, leading to iterative enhancements that optimize accuracy and predictive validity. This iterative process fosters a culture of innovation and continuous improvement, ultimately benefiting both educators and learners.

Third, evidence-based practice fosters transparency and accountability within the educational assessment landscape. By documenting the research methodologies, data sources, and validation procedures used to develop AI-powered assessment tools, researchers and practitioners can enhance the credibility and trustworthiness of their applications. This transparency not only instills confidence in different stakeholders (e.g., learners, educators, parents, and employers) but also encourages collaboration and knowledge-sharing within the broader educational community.

5.6 Democratizing AI in Education

AI technologies can immensely benefit education, offering new opportunities for teachers, students, and others involved. However, we must remain aware that these tools have the potential to amplify existing social inequalities. As AI-powered educational solutions become more widespread, it is essential from a fairness perspective to ensure that their advantages are accessible to all, irrespective of ethnicity, gender, disability status, or socioeconomic background. The three measures that ensure that AI benefits everyone include policies of accessibility testing, investing in open-source initiatives, and working more directly with measurement professionals and departments of education as representatives in the governance for the development of AI in the sector.

The divide between those with access to technology and those without access has been termed the ”Digital Divide” (Cullen, 2001). This issue manifests itself in many countries where rural and economically disadvantaged communities lack reliable internet service. As per Katz and Rideout (2021), in the United States of America, only 72 percent of Hispanic-headed households have broadband access compared to 80 percent of white families. This digital divide hinders digital literacy, which has already reduced employment opportunities, exacerbated disparities, and deepened social stratification. As AI-based tools primarily rely upon cloud-based services, the existing digital divide poses a real challenge in providing the benefits of AI-based tools to underprivileged students. The consequence is that the digital divide extends to AI-based skills that are becoming critical to future employment opportunities.

The integration of generative AI is increasingly being heralded as a huge business opportunity. However, it also can widen economic inequalities between those who possess the skills and resources to leverage AI and those who lack them. This divide is gaining recognition as the ”AI Divide” (Kitsara, 2022). In response, there is a growing discourse around the ”Democratization of AI” as a possible solution, which broadly calls for greater participation in the utilization, profitability, development, and governance of AI by all sections of society (Seger et al., 2023)). In the book ”AI for Everyone?: Critical Perspectives”, Verdegem (2021) argues that for AI to transform society positively, a ”radical democratization of AI” is essential, and this can be achieved by adhering to three fundamental principles:

- AI should be accessible to everyone. Nobody should be excluded from using AI because of differences in race, gender, class, or other distinctions.
- Developments in AI should contribute to the well-being of everyone in society.
- In a decent society, all members should have a say about what type of AI is being developed and what services are being offered.

From an educational standpoint, the first of these principles calls for developers of AI-based tools to work closely with educators and stakeholders to make AI accessible. As developers, this calls us to work closely to ensure disabilities do not adversely affect the usability of AI-powered tools. Policies concerning accessibility testing ensure that AI-based tools are usable to as many students as possible.

The primary obstacle to making AI accessible to everyone lies in closed-source models and data. As AI becomes more powerful and expensive, companies like OpenAI, Google, and Anthropic have sought to protect their intellectual property by restricting access to the model weights and training data. Not only does relying on closed-source solutions for building AI-based educational applications raise concerns regarding privacy and explainability (Gimpel et al., 2018), but it also means that the application would depend on a third party’s pricing structure. While current prices are quite reasonable, it is well-known that OpenAI operated at a significant loss of $540 million last year. If companies like OpenAI were to charge the total cost of running such large models, this move could make AI unaffordable for socioeconomically disadvantaged groups, thereby limiting the benefits of AI-based educational tools to only those who can afford the service.

Fortunately, some companies like Meta and Mistral have taken a positive step towards making AI more accessible by releasing open-source models such as Llama 3 and Mixtral (Jiang et al., 2024). These models have been benchmarked and evaluated, and their performance is competitive with closed-source models. Additionally, they have been released under licenses generally suitable for educational applications. Transitioning from closed-source initiatives towards open-source alternatives democratizes the governance and development of AI-based tools, enabling a broader range of people to contribute to the development of AI-based tools. It is also worth noting that libraries like the Transformer Library (Wolf et al., 2019) play a role in making AI more accessible by lowering the technical barriers to training and deploying AI in production environments. The availability of open-source models and software libraries that simplify the training and inference increases participation in developing AI-based tools, fostering collaboration. It creates a more competitive landscape that potentially reduces the cost of AI-based tools.

Critics who oppose the democratization of AI have raised significant concerns about democratizing the governance of AI systems (Himmelreich, 2023). They question how ethical standards for open-source models can be effectively imposed and legitimately governed by a majority. While these concerns primarily focus on the process of democratizing AI rather than whether it should be done, they highlight valid points. From an educational perspective, this discussion underscores the practical importance of involving measurement professionals and departments of education to serve as representatives for the interests of students and other stakeholders in the governance and development of AI-based educational tools. Their participation could help ensure that the equity and accessibility concerns of the educational sector are adequately addressed as AI becomes more democratized.

5.7 Environmental Impact of AI in Education

We are seeing LLMs being integrated into many aspects of educational technology. As noted in previous sections, these models are favored for their improvements in accuracy in the automated scoring of constructed response items. This accuracy comes at a significant increase in computational requirements, which, in turn, necessitates similar increases in carbon emissions (Strubell et al., 2020). It has been noted that training times for models like BERT are an order of magnitude longer than traditional n-gram approaches. However, the accuracy gains from transformer-based approaches can often be minimal (Mayfield & Black, 2020).

Regarding educational applications, evaluating whether adopting AI-based tools justifies the significant increase in costs and environmental emissions is crucial. In cases where these gains are significant, we need ways to mitigate the environmental impacts. We emphasize specific strategies discussed in Khowaja et al. (2024) that can help mitigate the environmental impact associated with the transition to LLMs. The first recommendation is to optimize computations. One way to do this is to pursue efficient model architectures. For example, the MobileBERT architecture is an example that uses bottlenecks to decrease the computational load imposed by the attention mechanism (Sun et al., 2020). Such architectures have been shown to perform comparably to much larger models in an educational context (Ormerod et al., 2021). For larger generative models, the adoption of parameter-efficient methods such as Quantized Low-Rank Adaptation has been estimated to reduce the overall carbon footprint of models by 72% (Dettmers et al., 2024). In addition to optimizing architectures, the parallelizable nature of transformer-based models means that they can be run efficiently on GPUs or specialized Tensor Processing Units (TPUs), which can reduce training times and overall electricity consumption (Khowaja et al., 2024).

The second recommendation is to ensure the energy used comes from renewable sources. If we are serving AI on local hardware, the percentage of energy coming from renewable sources depends only on the energy provider, so we only consider cloud computing services. Almost all cloud computing providers have demonstrated some commitment to renewable energy. We highlight the commitments made by the three cloud computing providers with the most significant market share :

- Amazon Web Services (31%): Through "The Climate Pledge," Amazon aims to achieve "Net-Zero Carbon" by 2040.
- Microsoft Azure (24%): Microsoft aims to have a negative footprint by 2030.
- Google Cloud Platform (11%): Google has been carbon neutral since 2007 and claims to be the first major company to achieve carbon neutrality.

This focus on the emissions from AI has prompted Lacoste et al. (2019) to develop a tool quantifying emissions by cloud computing provider, hardware, and region. This tool highlights that services like Google Cloud Provider and Azure generally seek to offset their emissions contribution, whereas others, such as AWS, do not. Ensuring AI-powered applications are powered by renewable energy or mitigated by suitable offsets can simply be a choice of appropriate cloud computing provider.

The third recommendation by Khowaja et al. (2024) is to encourage collaboration. Open-sourcing models, data, and collaborative research help foster innovation while reducing the duplication of efforts and resources. This being said, we acknowledge that there are practical barriers to some collaboration, such as data privacy and the application of intellectual property to software development in a corporate setting.

Educational measurement aims to leverage AI advancements to improve student outcomes. Despite the potential carbon footprint from AI-enabled educational tools being relatively small compared to overall emissions, no industry should be immune from the moral obligation to address climate change. We have an ethical responsibility to ensure that these tools are utilized responsibly and sustainably, necessitating a heightened awareness of the carbon footprint imposed by AI-based educational tools.

6 Discussion

Despite AI's potential to enhance educational measurement, researchers and practitioners must acknowledge and understand the limitations and ethical challenges associated with AI applications. An informed understanding will help prevent unquestioning belief in, over-reliance on, or misuse of AI technologies, especially in high-stakes settings where the implications and consequences of errors or biases can be significant. This paper outlined several ethical challenges common to many AI applications in educational assessment. First, AI technologies mirror and can even amplify biases in the training data; such models are often trained with unrepresentative data and thus inherently hold bias towards certain groups. Second, the transparency and explainability of AI technologies remain significant ethical challenges, particularly for tools that rely on third-party generative models. Third, AI technologies are often not evaluated before use, leading to issues such as creating items without good measurement properties or grading essays and assignments with low consistency. Fourth, the adoption of AI in educational assessments can influence broader social-environmental contexts, such as impacts on students' personal lives, equity of access to educational technologies for students from different backgrounds and cultures, and the environment.

Addressing these ethical challenges aligns with the principles outlined by Ammanath (2022), which emphasizes that AI tools for educational purposes must be thoughtfully designed, deployed, and monitored to ensure they are safe, robust, transparent, explainable, and responsible. Additionally, these ethical challenges are particularly important when viewed through the larger lens of the Standards (AERA, APA, NCME, 2014), which already contain a rigorous set of expectations for test takers, psychometricians, and test designers. Given that a new version of the Standards is currently under development, there may need to be an increased focus within the larger measurement community on the ethical use of AI. As the field increasingly adopts AI for educational measurement purposes, it is imperative to recognize when an AI tool may be unsafe, unfair, or unreliable. This vigilance will help harness AI's potential while safeguarding against its pitfalls, thereby advancing educational assessment in a responsible and ethical manner.

To successfully integrate AI technologies into the educational landscape, it is imperative to clearly define and attribute the roles played by human stakeholders and AI systems. By clearly delineating these roles, we can address the ethical challenge of accountability and use AI tools more judiciously. In the current paper, we argue that AI technologies should be applied to augment rather than replace human intelligence. In this framework, regardless of the aspect of educational assessments in which AI is employed (e.g., item generation, essay scoring, feedback generation, or proctoring), humans should maintain a supervisory role to ensure accuracy and responsible use. Therefore, humans must understand the reasons for AI systems' errors and biases and correct them. For humans to correct errors made by AI systems, it is essential to ensure that AI is understandable and transparent to its users (Shneiderman, 2022). For instance, in AIG tasks, users should understand how the characteristics of their prompts influence the outcomes from AI systems and how their choice of AI systems may impact the generated items, thereby effectively managing the quality and relevance of the generated content (Tan et al., 2024). Achieving this level of understanding and transparency may encourage increased collaboration between AI developers, measurement specialists, educators, educational researchers, and other stakeholders. By fostering such collaboration, the educational community can ensure that AI technologies are effectively, responsibly, and ethically integrated into the educational landscape.

As AI-powered tools are increasingly adopted within formal and informal education settings, it will require careful assurance that they are developed and adopted ethically, equitably, and effectively. This necessitates a dual-purpose, evidence-based educational measurement approach that scrutinizes the technical development of AI-powered tools and their real-world applications in educational contexts. The case of AWE systems serves as an excellent illustration of this need. On the development side, researchers must ensure that AWE systems are well-developed, usable, trustworthy, and unbiased. Furthermore, empirical research must analyze potential biases in training data, the outputs, and the validity and reliability of AI scoring (Goldshtein et al., 2024; Raczynski & Cohen, 2018; Wilson et al., 2019; Wind et al., 2018), while also examining the intended and unintended consequences (i.e., consequential validity) of using these AI measurement tools in authentic educational settings (Correnti et al., 2022, 2024; Wilson, Zhang, et al., 2024). Research exploring how AWE feedback is integrated into teaching practices and its effectiveness in enhancing student writing skills is especially important for ensuring that AI-driven AWE tools truly support educational objectives and do not inadvertently undermine critical human skills and relationships (Wilson et al., 2021). Thus, by embracing a holistic research approach that bridges technical and contextual analyses, the educational measurement community can better understand and leverage the potential of AI to enhance teaching and learning while mitigating any adverse or unintended consequences.

Similarly, it is essential to ensure clear, transparent, and effective communication with stakeholders when implementing AI in real-world educational assessments. These stakeholders often include students, parents, educators, policymakers, and other users of assessment scores or feedback. Such communication may include educating stakeholders about how the technology works, where and how it is used in the assessment process, how the results should or should not be interpreted, and addressing concerns and potential misconceptions. In this way, stakeholders can develop a realistic understanding of the AI's capabilities and limitations, which is crucial for building trust and acceptance. Additionally, open dialogues about the ethical considerations and safeguards in place to protect the integrity of the assessment process could help to further reinforce confidence among stakeholders and encourage test-takers to give their best performance.

Adopting new technologies like AI in the educational sector must also be done through an ethical lens that prioritizes environmental sustainability. No industry can afford to be insulated from the moral imperative to confront climate change head-on. Although the deployment of generative AI carries a larger carbon footprint than traditional approaches, this paper has outlined several pragmatic strategies to mitigate the climate impacts stemming from generative AI usage. Key recommendations include the proliferation of open-source AI models that can be leveraged and improved collaboratively, reducing redundancies. Continued research into developing computationally efficient AI methods that require less energy-intensive training is also crucial. Moreover, educational institutions should aim to harness cloud computing services that are carbon neutral through renewable energy offsets or direct procurement of clean energy.

6.1 Directions for Future Research

There are several directions for future AI and educational measurement work, especially around fairness, environmental impact, and explainability. The first of these is creating a set of evidence-based metrics and benchmarks to ensure that AI-based systems treat all students fairly, regardless of demographic background, with respect to educational tasks such as grading, giving feedback, and making predictions around student outcomes. Another option to improve model fairness may be researching and creating a series of vetted training datasets that can be used as additional training material to represent populations who have historically tended to experience AI bias. Additional fairness work may also include research into different human-in-the-loop configurations with respect to high-stakes applications of AI in education such that the AI system alone is not making the final decision when the outcome of these decisions may have a lasting impact on the student.

Second, there is often a large fiscal and energy burden to training brand-new AI systems as well as the ongoing maintenance and use of these systems Khowaja et al. (2024). More research needs to be performed, such as that of Dettmers et al. (2024), to find new ways to design these AI models to be more efficient and thus lower the energy burden of these systems. An additional approach may be to expand research on optimizing and fine-tuning existing general LLMs for specific educational purposes instead of incurring the additional energy burden of training new models.

Third, to meet the current rigorous expectations of “Rights and Responsibilities of Test Takers” found in the Standards (AERA, APA, NCME, 2014), more work is needed to ensure that there are rigorous standards for AI model explainability that would allow any test taker to inquire how their score was calculated. Currently, one of the main concerns over the use of LLM-based tools for scoring purposes is that these models contain thousands, millions, or billions of parameters that make it very difficult for humans to understand all of the aspects of how LLM output was derived; we have referred to this above as the "black box" problem. In fact, not even experts and LLM developers can reliably interpret how a specific output was reached (Bowman, 2023). To encourage deeper trust in using these tools for high-stakes purposes, more research must be performed to foster a deeper understanding of and dissemination about how these scores are generated to support test takers in knowing more about how their scores are being constructed.

Furthermore, while we provide robust and representative examples of assessments in this paper, it is crucial to recognize the ongoing need to expand our data sources continually. We emphasize the importance of including AI-powered assessment tools designed for culturally and linguistically diverse populations. We encourage future research to incorporate a broader range of these tools from various perspectives and contexts. This strategy will help address potential biases and limitations in data collection, ensuring that AI tools are equitable and effective for all users. By diversifying the data to include culturally and linguistically relevant tools, we can enhance the accuracy and fairness of AI applications in education, leading to more inclusive and reliable outcomes.

6.2 Conclusion

This paper has explored broader issues surrounding fairness and equity in AI. We have underscored the critical need to democratize AI technologies, ensuring they remain widely accessible resources that can benefit all segments of society equitably. This democratization entails more than just making AI available–it necessitates proactive efforts to guarantee that the development and implementation of these powerful technologies occur through an inclusive process that empowers diverse voices and perspectives. There is a genuine danger that advanced AI capabilities could become concentrated in the hands of a privileged few nations, corporations, or elite groups. True democratization in education demands that we foster ecosystems where AI development is a participatory endeavor, drawing from the experience of teachers, students, and assessment professionals within government and industry. We can only shape AI to be genuinely universally beneficial through multi-stakeholder collaboration. Seeking the best way to foster such ecosystems requires careful consideration and ensuring that our AI systems align with and abide by human values.

References

AERA, APA, NCME. (2014). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Almond, P., Winter, P., Cameto, R., Russell, M., Sato, E., Clarke-Midura, J., … Lazarus, S. (2010). Technology-enabled and universally designed assessment: Considering access in measuring the achievement of students with disabilities—A foundation for research. *The Journal of Technology, Learning and Assessment*, *10*(5). (Number: 5)

Alsabhan, W. (2023). Student cheating detection in higher education by implementing machine learning and lstm techniques. *Sensors*, *23*(8), 4149. doi: 10.3390/s23084149

Alwahaby, H., Cukurova, M., Papamitsiou, Z., & Giannakos, M. (2022). The evidence of impact and ethical considerations of multimodal learning analytics: A systematic literature review. In M. Giannakos, D. Spikol, D. Di Mitri, K. Sharma, X. Ochoa, & R. Hammad (Eds.), *The multimodal learning analytics handbook* (pp. 289–325). Cham: Springer International Publishing. doi: 10.1007/978-3-031-08076-0_12

Ammanath, B. (2022). *Trustworthy AI: A business guide for navigating trust and ethics in AI.* John Wiley & Sons.

Andersen, Ø. E., Yuan, Z., Watson, R., & Cheung, K. Y. F. (2021). *Benefits of alternative evaluation methods for automated essay scoring* (Tech. Rep.). Paris, France: International Educational Data Mining Society.

Angoff, W. H. (2012). Perspectives on differential item functioning methodology. In *Differential item functioning* (pp. 3–23). Routledge.

Arora, P. (2021). *Is remote proctoring the future of academia?* Retrieved from https://elearningindustry.com/is-remote-proctoring-future-academia

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/3442188.3445922

Bergmans, L., Bouali, N., Luttikhuis, M., & Rensink, A. (2021). On the efficacy of online proctoring using Proctorio. In *Proceedings of the 13th International Conference on Computer Supported Education (CSEDU 2021)* (pp. 279–290). SCITEPRESS. doi: 10.5220/0010399602790290

Biecek, P., & Burzykowski, T. (2021). *Explanatory model analysis: Explore, explain, and examine predictive models*. Chapman and Hall/CRC.

Bilen, E., & Matros, A. (2021). Online cheating amid COVID-19. *Journal of Economic Behavior & Organization*, *182*, 196–211. doi: 10.1016/j.jebo.2020.12.004

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, *21*(1), 5–31. doi: 10.1007/s11092-008-9068-5

Blikstein, P., & Worsley, M. (2016). Multimodal learning analytics and education data mining: Using computational technologies to measure complex learning tasks. *Journal of Learning Analytics*, *3*(2), 220–238. doi: 10.18608/jla.2016.32.11

Bond, R. R., Mulvenna, M. D., Wan, H., Finlay, D. D., Wong, A., Koene, A., … Adel, T. (2019). Human centered artificial intelligence: Weaving UX into algorithmic decision making. In *Rochi* (pp. 2–9).

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

Boulanger, D., & Kumar, V. S. (2024). Explainable AI and AWE: Balancing tensions between transparency and predictive accuracy. In *The Routledge international handbook of automated essay evaluation* (pp. 445–468). Routledge.

Bowman, S. R. (2023). *Eight things to know about large language models.*

Bozkurt, A., & Sharma, R. C. (2023). Generative AI and prompt engineering: The art of whispering to let the genie out of the algorithmic world. *Asian Journal of Distance Education*, *18*(2), i–vii. Retrieved from https://www.asianjde.com/ojs/index.php/AsianJDE/article/view/749

Breyer, F. J., Rupp, A. A., & Bridgeman, B. (2017). Implementing a contributory scoring approach for the GRE® analytical writing section: A comprehensive empirical investigation. *ETS Research Report Series*, *2017*(1), 1–28.

Bulut, O., Gorgun, G., & Karamese, H. (2023). Incorporating test-taking engagement into multistage adaptive testing design for large-scale assessments. *Journal of Educational Measurement*. doi: 10.1111/jedm.12380

Bulut, O., Gorgun, G., Yildirim-Erbasli, S. N., Wongvorachan, T., Daniels, L. M., Gao, Y., … Shin, J. (2023). Standing on the shoulders of giants: Online formative assessments as the foundation for predictive learning analytics models. *British Journal of Educational Technology*, *54*(1), 19-39. doi: https://doi.org/10.1111/bjet.13276

Bulut, O., Wongvorachan, T., He, S., & Lee, S. (2024). *Enhancing high-school dropout identification: A collaborative approach integrating human and machine insights.* doi: 10.21203/rs.3.rs-3871667/v1

Bulut, O., & Yildirim-Erbasli, S. N. (2022). Automatic story and item generation for reading comprehension assessments with transformers. *International Journal of Assessment Tools in Education*, *9*(Special Issue), 72–87. doi: 10.21449/ijate.1124382

Bulut, O., Yildirim-Erbasli, S. N., & Gorgun, G. (2024). Assessment analytics for digital assessments identifying, modeling, and interpreting behavioral engagement. In M. Sahin & D. Ifenthaler (Eds.), *Assessment analytics in education: Designs, methods and solutions* (pp. 35–60). Cham: Springer International Publishing. doi: 10.1007/978-3-031-56365-2_3

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (pp. 77–91).

Burstein, J., Elliot, N., Klebanov, B. B., Madnani, N., Napolitano, D., Schwartz, M., … Molloy, H. (2018). Writing MentorTM: Writing progress using self-regulated writing support. *The Journal of Writing Analytics*, *2*(1), 285–313. doi: 10.37514/JWA-J.2018.2.1.12

Burstein, J., Kukich, K., Wolff, S., Lu, C., Chodorow, M., Braden-Harder, L., & Harris, M. D. (1998). Automated scoring using a hybrid feature identification technique. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1* (pp. 206–210). USA: Association for Computational Linguistics. doi: 10.3115/980845.980879

Burstein, J., von Davier, A., Yancey, K., Belzak, W., Bicknell, K., Gottlieb, C., … Zheng, M. (2023). *DET responsible AI standards 2024.* Retrieved from https://duolingo-papers.s3.amazonaws.com/

other/DET%2BResponsible%2BAI%2BStandards%2B-%2B040824.pdf

Buzick, H., Oliveri, M. E., Attali, Y., & Flor, M. (2016). Comparing human and automated essay scoring for prospective graduate students with learning disabilities and/or ADHD. *Applied Measurement in Education*, *29*(3), 161–172. doi: 10.1080/08957347.2016.1171765

Cairns, A. W., Bond, R. R., Finlay, D. D., Breen, C., Guldenring, D., Gaffney, R., . . . Henn, P. (2016). A computer-human interaction model to improve the diagnostic accuracy and clinical decision-making during 12-lead electrocardiogram interpretation. *Journal of Biomedical Informatics*, *64*, 93–107. doi: 10.1016/j.jbi.2016.09.016

Carless, D. (2019). Feedback loops and the longer-term: Towards feedback spirals. *Assessment & Evaluation in Higher Education*, *44*(5), 705–714. doi: 10.1080/02602938.2018.1531108

Casuat, C. D., & Festijo, E. D. (2019, December). Predicting students' employability using machine learning approach. In *2019 IEEE 6th International Conference on Engineering Technologies and Applied Sciences (ICETAS)* (pp. 1–5). doi: 10.1109/ICETAS48360.2019.9117338

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority oversampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. doi: 10.1613/jair.953

Chen, C.-F. E., & Cheng, W.-Y. E. (2008). Beyond the design of automated writing evaluation: Pedagogical practices and perceived learning effectiveness in EFL writing classes. *Language Learning & Technology*, *12*(2), 94–112. doi: 10125/44145

Chen, G., Rolim, V., Mello, R. F., & Gašević, D. (2020, March). Let's shine together! a comparative study between learning analytics and educational data mining. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge* (pp. 544–553). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/3375462.3375500

Choi, I., & Johnson, M. S. (2024, April). Examining partial derivatives to identify causes of differential prediction bias in automated scores. *Paper presented at the annual meeting of the National Council for Measurement in Education, Philadelphia, PA.*

Cizek, G. J., & Wollack, J. A. (2016). *Handbook of quantitative methods for detecting cheating on tests.* Taylor & Francis.

Coghlan, S., Miller, T., & Paterson, J. (2021). Good proctor or "big brother"? Ethics of online exam supervision technologies. *Philosophy & Technology*, *34*(4), 1581–1606. doi: 10.1007/s13347-021-00476-1

Correnti, R., Matsumura, L. C., Wang, E. L., Litman, D., & Zhang, H. (2022). Building a validity argument for an automated writing evaluation system (erevise) as a formative assessment. *Computers and Education Open*, *3*, 100084. doi: 10.1016/j.caeo.2022.100084

Correnti, R., Wang, E. L., Matsumura, L. C., Litman, D., Liu, Z., & Li, T. (2024). Supporting students' text-based evidence use via formative automated writing and revision assessment. In *The Routledge International Handbook of Automated Essay Evaluation* (pp. 221–243). Routledge.

Cotos, E. (2023). Automated feedback on writing. In O. Kruse et al. (Eds.), *Digital writing technologies in higher education: Theory, research, and practice* (pp. 347–364). Cham: Springer International Publishing. doi: 10.1007/978-3-031-36033-6_22

Cullen, R. (2001). Addressing the digital divide. *Online information review*, *25*(5), 311–320.

Dawson, P. (2024). Remote proctoring: Understanding the debate. In S. E. Eaton (Ed.), *Second handbook of academic integrity* (pp. 1511–1526). Cham: Springer Nature Switzerland. doi: 10.1007/978-3-031-54144-5_150

Deeva, G., Bogdanova, D., Serral, E., Snoeck, M., & De Weerdt, J. (2021). A review of automated feedback systems for learners: Classification framework, challenges and opportunities. *Computers & Education*, *162*, 104094. doi: 10.1016/j.compedu.2020.104094

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2024). Qlora: Efficient finetuning of quantized LLMs. *Advances in Neural Information Processing Systems*, *36*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding.* arXiv. (arXiv:1810.04805 [cs]) doi: 10.48550/arXiv.1810.04805

Drori, I., Zhang, S., Shuttleworth, R., Tang, L., Lu, A., Ke, E., ... Strang, G. (2022). A neural network solves, explains, and generates university math problems by program synthesis and few-shot learning at human level. *Proceedings of the National Academy of Sciences*, *119*(32), e2123433119. doi: 10.1073/

pnas.2123433119

Dyer, J. (2024). Framework for ethical implementation of remote proctoring in education. In S. E. Eaton (Ed.), *Second handbook of academic integrity* (pp. 1527–1550). Cham: Springer Nature Switzerland. doi: 10.1007/978-3-031-54144-5_151

European Commission. (2019). *The European Commission's high-level expert group on intelligence: Ethics guidelines for trustworthy AI* (Tech. Rep.). Publications Office of the European Union. Retrieved from https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai

European Commission, Directorate-General for Education, Youth, S., & Culture. (2022). *Ethical guidelines on the use of artificial intelligence (AI) and data in teaching and learning for educators* (Tech. Rep.). Publications Office of the European Union. Retrieved from https://data.europa.eu/doi/10.2766/153756

Fleckenstein, J., Liebenow, L. W., & Meyer, J. (2023). Automated feedback and writing: A multi-level meta-analysis of effects on students' performance. *Frontiers in Artificial Intelligence*, *6*, 1162454.

Flor, M., & Hao, J. (2021). Text mining and automated scoring. In A. A. von Davier, R. J. Mislevy, & J. Hao (Eds.), *Computational psychometrics: New methodologies for a new generation of digital learning and assessment: With examples in R and Python* (pp. 245–262). Cham: Springer International Publishing. doi: 10.1007/978-3-030-74394-9_14

Fritts, M., & Cabrera, F. (2021). AI recruitment algorithms and the dehumanization problem. *Ethics and Information Technology*, *23*(4), 791–801. doi: 10.1007/s10676-021-09615-w

Fu, Q.-K., Zou, D., Xie, H., & Cheng, G. (2024). A review of AWE feedback: types, learning outcomes, and implications. *Computer Assisted Language Learning*, *37*(1-2), 179–221. doi: 10.1080/09588221.2022.2033787

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., . . . Ahmed, N. K. (2024). *Bias and fairness in large language models: A survey.* arXiv. doi: 10.48550/arXiv.2309.00770

Gierl, M. J., & Haladyna, T. M. (2012). Automatic item generation: An introduction. In *Automatic item generation.* Routledge.

Gierl, M. J., & Lai, H. (2012). The role of item models in automatic item generation. *International Journal* *of Testing*, *12*(3), 273–298. doi: 10.1080/15305058.2011.635830

Gierl, M. J., Shin, J., Firoozi, T., & Lai, H. (2022). Using content coding and automatic item generation to improve test security. *Frontiers in Education*, *7*, 853578. doi: 10.3389/feduc.2022.853578

Gimpel, H., Kleindienst, D., Nüske, N., Rau, D., & Schmied, F. (2018). The upside of data privacy – delighting customers by implementing data privacy measures. *Electronic Markets*, *28*(4), 437–452. doi: 10.1007/s12525-018-0296-3

Goldshtein, M., Alhashim, A. G., & Roscoe, R. D. (2024). Automating bias in writing evaluation: Sources, barriers, and recommendations. In *The Routledge international handbook of automated essay evaluation* (pp. 421–444). Routledge.

Gorgun, G., & Bulut, O. (2021). A polytomous scoring approach to handle not-reached items in low-stakes assessments. *Educational and Psychological Measurement*, *81*(5), 847–871. doi: 10.1177/0013164421991211

Gorgun, G., & Bulut, O. (2022). Identifying aberrant responses in intelligent tutoring systems: an application of anomaly detection methods. *Psychological Test and Assessment Modeling*, *64*(4), 359–384.

Gorgun, G., & Bulut, O. (2023). Incorporating test-taking engagement into the item selection algorithm in low-stakes computerized adaptive tests. *Large-scale Assessments in Education*, *11*(1), 27. doi: 10.1186/s40536-023-00177-5

Grimes, D., & Warschauer, M. (2010). Utility in a fallible tool: A multi-site case study of automated writing evaluation. *The Journal of Technology, Learning and Assessment*, *8*(6).

Haller, S., Aldea, A., Seifert, C., & Strisciuglio, N. (2022). *Survey on automated short answer grading with deep learning: From word embeddings to transformers.* arXiv. doi: 10.48550/arXiv.2204.03503

Hao, J., von Davier, A. A., Yaneva, V., Lottridge, S., von Davier, M., & Harris, D. J. (2024). Transforming assessment: The impacts and implications of large language models and generative AI. *Educational Measurement: Issues and Practice*, *43*(2), 16–29. doi: 10.1111/emip.12602

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

He, Y., Jing, S., & Lu, Y. (2022, April). A multilevel multinomial logit approach to bias detection. In *the* *annual meeting of the National Council on Measurement in Education. San Diego, CA.*

Heston, T. F., & Khun, C. (2023). Prompt engineering in medical education. *International Medical Education*, *2*(3), 198–205. doi: 10.3390/ime2030019

Higgins, D., Xi, X., Zechner, K., & Williamson, D. (2011). A three-stage approach to the automated scoring of spontaneous spoken responses. *Computer Speech & Language*, *25*(2), 282–306. doi: 10.1016/j.csl.2010.06.001

Himmelreich, J. (2023). Against "democratizing AI". *AI & Society*, *38*(4), 1333–1346. doi: 10.1007/s00146-021-01357-z

Hockly, N. (2019). Automated writing evaluation. *ELT Journal*, *73*(1), 82–88. doi: 10.1093/elt/ccy044

Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning.* Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Holling, H., Bertling, J. P., & Zeuch, N. (2009). Automatic item generation of probability word problems. *Studies in Educational Evaluation*, *35*(2), 71–76. doi: 10.1016/j.stueduc.2009.10.004

Huawei, S., & Aryadoust, V. (2023). A systematic review of automated writing evaluation systems. *Education and Information Technologies*, *28*(1), 771–795. doi: 10.1007/s10639-022-11200-7

International Test Commission. (2014). ITC guidelines on quality control in scoring, test analysis, and reporting of test scores. *International Journal of Testing*, *14*(3), 195–217. doi: 10.1080/15305058.2014.918040

ITC, & ATP. (2022). *The guidelines for technology-based assessment.* Association of Test Publishers, International Test Commission. Retrieved from https://www.intestcom.org/page/16

Jaschik, S. (2021). *ProctorU abandons business based solely on AI.* Retrieved from https://www.insidehighered.com/news/2021/05/24/proctoru-abandons-business-based-solely-ai

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., … others (2024). Mixtral of experts. *arXiv Preprint*. doi: 10.48550/arXiv.2401.04088

Jiao, H., Yadav, C., & Li, G. (2023). *Integrating psychometric analysis and machine learning to augment data for cheating detection in large-scale assessment.* OSF. doi: 10.31234/osf.io/fjz2c

Jin, S. (2012). Design of an online learning platform with Moodle. In *the 7th International Conference on*

Computer Science & Education (ICCSE) (pp. 1710–1714). doi: 10.1109/ICCSE.2012.6295395

Johnson, M. S., & McCaffrey, D. F. (2023). Evaluating fairness of automated scoring in educational measurement. *Advancing Natural Language Processing in Educational Assessment*, 142.

Jones, B. D., & Egley, R. J. (2007). Learning to take tests or learning for understanding? teachers' beliefs about test-based accountability. *The Educational Forum*, *71*(3), 232–248. doi: 10.1080/00131720709335008

Jones-Jang, S. M., & Park, Y. J. (2023). How do people react to AI failure? automation bias, algorithmic aversion, and perceived controllability. *Journal of Computer-Mediated Communication*, *28*(1). doi: 10.1093/jcmc/zmac029

Justice, D. (2022, April). A linear model approach to bias detection. In *the annual meeting of the National Council on Measurement in Education.* San Diego, CA.

Kamalov, F., Sulieman, H., & Calonge, D. S. (2021). Machine learning based approach to exam cheating detection. *PLOS ONE*, *16*(8), e0254340. doi: 10.1371/journal.pone.0254340

Katz, V., & Rideout, V. (2021). Learning at home while under-connected: Lower-income families during the COVID-19 pandemic. *New America*. Retrieved from https://www.carnegie.org/publications/learning-home-while-under-connected-lower-income-families-during-covid-19-pandemic/

Ke, Z., & Ng, V. (2019, August). Automated essay scoring: A survey of the state of the art. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence* (pp. 6300–6308). Macao, China. doi: 10.24963/ijcai.2019/879

Khera, R., Simon, M. A., & Ross, J. S. (2023). Automation bias and assistive AI: Risk of harm from AI-driven clinical decision support. *JAMA*, *330*(23), 2255–2257. doi: 10.1001/jama.2023.22557

Khowaja, S. A., Khuwaja, P., Dev, K., Wang, W., & Nkenyereye, L. (2024). Chatgpt needs spade (sustainability, privacy, digital divide, and ethics) evaluation: A review. *Cognitive Computation*, 1–23. doi: 10.1007/s12559-024-10285-1

Kim, D., Woo, A., & Dickison, P. (2016). Identifying and investigating aberrant responses using psychometrics-based and machine learning-based approaches. In G. J. Cizek & J. A. Wollack (Eds.), *Handbook of quantitative methods for detecting cheating on tests* (pp. 70–97). Routledge.

Kitsara, I. (2022). Artificial intelligence and the digital divide: From an innovation perspective. In *Platforms and artificial intelligence: The next generation of competences* (pp. 245–265). Springer.

Kumari, V., Keshari, S., Sharma, Y., & Goel, L. (2022). Context-based question answering system with suggested questions. In *the 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence)* (pp. 368–373). Noida, India.

Kupfer, C., Prassl, R., Fleiß, J., Malin, C., Thalmann, S., & Kubicek, B. (2023). Check the box! how to deal with automation bias in ai-based personnel selection. *Frontiers in Psychology*, *14*, 1118723. doi: 10.3389/fpsyg.2023.1118723

Lacoste, A., Luccioni, A., Schmidt, V., & Dandres, T. (2019). Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*. doi: 10.48550/arXiv.1910.09700

Lai, H., Alves, C., & Gierl, M. (2009, June). Using automatic item generation to address item demands for CAT. In *the 2009 GMAC Conference on Computerized Adaptive Testing.* Retrieved from `http://iacat.org/sites/default/files/biblio/cat09lai.pdf`

Lee, U., Jung, H., Jeon, Y., Sohn, Y., Hwang, W., Moon, J., & Kim, H. (2023). Few-shot is enough: Exploring ChatGPT prompt engineering method for automatic question generation in english education. *Education and Information Technologies*. doi: 10.1007/s10639-023-12249-8

Li, J., Tang, T., Zhao, W. X., Nie, J.-Y., & Wen, J.-R. (2024). Pre-trained language models for text generation: A survey. *ACM Computing Surveys*, *56*(9), 1–39. doi: 10.1145/3649449

Li, R. (2023). Still a fallible tool? Revisiting effects of automated writing evaluation from activity theory perspective. *British Journal of Educational Technology*, *54*(3), 773–789. doi: 10.1111/bjet.13294

Lin, B.-R., & Kifer, D. (2015). Information measures in statistical privacy and data processing applications. *ACM Transactions on Knowledge Discovery from Data*, *9*(4), 28:1–28:29. doi: 10.1145/2700407

Lindner, M. A., Lüdtke, O., & Nagy, G. (2019). The onset of rapid-guessing behavior over the course of testing time: A matter of motivation and cognitive resources. *Frontiers in Psychology*, *10*. doi: 10.3389/fpsyg.2019.01533

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, *55*(9), 1–35. doi: 10.1145/3560815

Liu, X., & Fauss, M. (2024, April). A Bayesian nonparametric model for flexible automated scoring. *Paper presented at the annual meeting of the National Council for Measurement in Education, Philadelphia, PA.*

Liu, Y., Cheng, Y., & Liu, H. (2020). Identifying effortful individuals with mixture modeling response accuracy and response time simultaneously to improve item parameter estimation. *Educational and Psychological Measurement*, *80*(4), 775–807. doi: 10.1177/0013164419895068

Lottridge, S., Ormerod, C., & Jafari, A. (2023). Psychometric considerations when using deep learning for automated scoring. In V. Yaneva & M. von Davier (Eds.), *Advancing natural language processing in educational assessment* (pp. 15–30). Routledge. doi: 10.4324/9781003278658-3

Lottridge, S., & Young, M. (2022, April). Examining bias in automated scoring of reading comprehension items. In *the annual meeting of the National Council on Measurement in Education.* San Diego, CA.

Lyell, D., & Coiera, E. (2017). Automation bias and verification complexity: A systematic review. *Journal of the American Medical Informatics Association*, *24*(2), 423–431. doi: 10.1093/jamia/ocw105

Mayfield, E., & Black, A. W. (2020). Should you fine-tune BERT for automated essay scoring? In *Proceedings of the fifteenth workshop on innovative use of NLP for building educational applications* (pp. 151–162).

McCaffrey, D. F., Casabianca, M., J., & Johnson, M. S. (2024). *The proportional reduction in mean squared error for use in automated scoring.* (Unpublished manuscript)

McCaffrey, D. F., Casabianca, J. M., Ricker-Pedley, K. L., Lawless, R. R., & Wendler, C. (2022). Best practices for constructed-response scoring. *ETS Research Report Series*, *2022*(1), 1–58.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, *54*(6), 115:1–115:35. doi: 10.1145/3457607

Meng, H., & Ma, Y. (2023). Machine learning–based profiling in test cheating detection. *Educational Measurement: Issues and Practice*, *42*(1), 59–75. doi: 10.1111/emip.12541

Molnar, C. (2020). *Interpretable machine learning: A guide for making black box models explainable.* Retrieved from https://christophm.github.io/interpretable-ml-book/

Moore, N. S., & MacArthur, C. A. (2016). Student use of automated essay evaluation technology during revision. *Journal of Writing Research*, *8*(1), 149–175. doi: 10.17239/jowr-2016.08.01.05

Moreno-Guerrero, A.-J., Rodríguez-Jiménez, C., Gómez-García, G., & Ramos Navas-Parejo, M. (2020). Educational innovation in higher education: Use of role playing and educational video in future teachers' training. *Sustainability*, *12*(6), 2558. doi: 10.3390/su12062558

Motwani, S., Nagpal, C., Motwani, M., Nagdev, N., & Yeole, A. (2021). *AI-based proctoring system for online tests* [SSRN Scholarly Paper]. Rochester, NY. Retrieved from `https://papers.ssrn.com/abstract=3866446` doi: 10.2139/ssrn.3866446

Mueller, L., Zhang, Y., & Ferrara, S. (2016). What have we learned? In G. J. Cizek & J. A. Wollack (Eds.), *Handbook of quantitative methods for detecting cheating on tests* (pp. 373–389). Routledge.

Myers, M. C., & Wilson, J. (2023). Evaluating the construct validity of an automated writing evaluation system with a randomization algorithm. *International Journal of Artificial Intelligence in Education*, *33*(3), 609–634. doi: 10.1007/s40593-022-00301-6

Ngo, T. T.-N., Chen, H. H.-J., & Lai, K. K.-W. (2024). The effectiveness of automated writing evaluation in EFL/ESL writing: a three-level meta-analysis. *Interactive Learning Environments*, *32*(2), 727–744. doi: 10.1080/10494820.2022.2096642

Nguyen, A., Ngo, H. N., Hong, Y., Dang, B., & Nguyen, B.-P. T. (2023). Ethical principles for artificial intelligence in education. *Education and Information Technologies*, *28*(4), 4221–4241. doi: 10.1007/s10639-022-11316-w

Nigam, A., Pasricha, R., Singh, T., & Churi, P. (2021). A systematic review on AI-based proctoring systems: Past, present and future. *Education and Information Technologies*, *26*(5), 6421–6445. doi: 10.1007/s10639-021-10597-x

Nunes, A., Cordeiro, C., Limpo, T., & Castro, S. L. (2022). Effectiveness of automated writing evaluation systems in school settings: A systematic review of studies from 2000 to 2020. *Journal of Computer Assisted Learning*, *38*(2), 599–620. doi: 10.1111/jcal.12635

OECD. (2023). Opportunities, guidelines and guardrails for effective and equitable use of AI in education. In *OECD Digital Education Outlook 2023: Towards an Effective Digital Education Ecosystem.* OECD Publishing. Retrieved from `https://doi.org/10.1787/2b39e98b-en.`

Offerijns, J., Verberne, S., & Verhoef, T. (2020). Better distractions: Transformer-based distractor generation and multiple choice question filtering. *arXiv preprint arXiv:2010.09598*.

Ormerod, C. M. (2022a). Mapping between hidden states and features to validate automated essay scoring using DeBERTa models. *Psychological Test and Assessment Modeling*, *64*(4), 495–526.

Ormerod, C. M. (2022b). Short-answer scoring with ensembles of pretrained language models. *arXiv Preprint*. doi: 10.48550/arXiv.2202.11558

Ormerod, C. M., Malhotra, A., & Jafari, A. (2021). Automated essay scoring using efficient transformer-based language models. *arXiv preprint*. doi: 10.48550/arXiv.2102.13136

Page, E. B. (1966). The imminence of... grading essays by computer. *The Phi Delta Kappan*, *47*(5), 238–243. Retrieved from `https://www.jstor.org/stable/20371545`

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, *39*(2), 230–253. doi: 10.1518/001872097778543886

Penfield, R. D. (2016). Fairness in test scoring. In N. J. Dorans & L. L. Cook (Eds.), *Fairness in educational assessment and measurement* (pp. 55–75). Routledge.

Raczynski, K., & Cohen, A. (2018). Appraising the scoring performance of automated essay scoring systems—Some additional considerations: Which essays? Which human raters? Which scores? *Applied Measurement in Education*, *31*, 233–240. doi: 10.1080/08957347.2018.1464449

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training.* Retrieved from `https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf`

Rahm, L. (2023). Education, automation and AI: a genealogy of alternative futures. *Learning, Media and Technology*, *48*(1), 6–24. doi: 10.1080/17439884.2021.1977948

Ramesh, D., & Sanampudi, S. K. (2022). An automated essay scoring systems: A systematic literature review. *Artificial Intelligence Review*, *55*(3), 2495–2527. doi: 10.1007/s10462-021-10068-2

Ranger, J., Schmidt, N., & Wolgast, A. (2023). Detecting cheating in large-scale assessment: The transfer of detectors to new tests. *Educational and Psychological Measurement*, *83*(5), 1033–1058. doi: 10.1177/00131644221132723

Rao, K. (2015). Universal design for learning and multimedia technology: Supporting culturally and linguistically diverse students. *Journal of Educational Multimedia and Hypermedia*, *24*(2), 121–137. Retrieved from http://hdl.handle.net/10125/41065

Reiss, M. J. (2021). The use of Al in education: Practicalities and ethical considerations. *London Review of Education*, *19*(1).

Riordan, B., Bichler, S., Bradford, A., King Chen, J., Wiley, K., Gerard, L., & C. Linn, M. (2020, July). An empirical investigation of neural methods for content scoring of science explanations. In J. Burstein et al. (Eds.), *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 135–144). Seattle, WA, USA: Association for Computational Linguistics. doi: 10.18653/v1/2020.bea-1.13

Riordan, B., Bichler, S., Bradford, A., & Linn, M. C. (2020). Probing saliency in short answer scoring models for science explanations. *New York Academy of Sciences Natural Language, Dialog and Speech Symposium*.

Rodriguez-Torrealba, R., Garcia-Lopez, E., & Garcia-Cabot, A. (2022). End-to-end generation of multiple-choice questions using text-to-text transfer transformer models. *Expert Systems with Applications*, *208*, 118258. doi: 10.1016/j.eswa.2022.118258

Sankey, M., Birch, D., & Gardiner, M. W. (2010). Engaging students through multimodal learning environments: The journey continues. *Proceedings of the 27th Australasian Society for Computers in Learning in Tertiary Education*, 852–863.

Säuberli, A., & Clematide, S. (2024). Automatic generation and evaluation of reading comprehension test items with large language models. *arXiv preprint arXiv:2404.07720*.

Sayin, A., & Gierl, M. (2024). Using OpenAI GPT to generate reading comprehension items. *Educational Measurement: Issues and Practice*, *43*(1), 5–18. doi: 10.1111/emip.12590

Seger, E., Ovadya, A., Siddarth, D., Garfinkel, B., & Dafoe, A. (2023). Democratising ai: Multiple meanings, goals, and methods. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 715–722).

Sharma, K., & Giannakos, M. (2020). Multimodal data capabilities for learning: What can multimodal data tell us about learning? *British Journal of Educational Technology*, *51*(5), 1450-1484. doi: https://doi.org/10.1111/bjet.12993

Shermis, M. D. (2024). AI scoring and writing fairness. In *The Routledge international handbook of automated essay evaluation* (pp. 386–420). Routledge.

Shermis, M. D., Mao, L., Mulholland, M., & Kieftenbeld, V. (2017). Use of automated scoring features to generate hypotheses regarding language-based DIF. *International Journal of Testing*, *17*(4), 351–371. doi: 10.1080/15305058.2017.1308949

Shi, Z., Liu, F., Lai, C., & Jin, T. (2022). Enhancing the use of evidence in argumentative writing through collaborative processing of content-based automated writing evaluation feedback. *Language Learning & Technology*. (Publisher: University of Hawaii, National Foreign Language Resource Center)

Shneiderman, B. (2022). *Human-centered AI.* Oxford University Press.

Singh, J., Aggarwal, R., Tiwari, S., & Joshi, V. (2022, October). Exam proctoring classification using eye gaze detection. In *the 3rd International Conference on Smart Electronics and Communication* (pp. 371–376). doi: 10.1109/ICOSEC54921.2022.9951987

Slusky, L. (2020). Cybersecurity of online proctoring systems. *Journal of International Technology and Information Management*, *29*(1), 56–83. doi: 10.58729/1941-6679.1445

Smith, B. E., Pacheco, M. B., & Khorosheva, M. (2021). Emergent bilingual students and digital multimodal composition: A systematic review of research in secondary classrooms. *Reading Research Quarterly*, *56*(1), 33–52. doi: 10.1002/rrq.298

Soland, J., Kuhfeld, M., & Rios, J. (2021). Comparing different response time threshold setting methods to detect low effort on a large-scale assessment. *Large-scale Assessments in Education*, *9*(1), 8. doi: 10.1186/s40536-021-00100-w

Stahl, W. M., & Karger, J. (2016). Student data privacy, digital learning, and special education: Challenges at the intersection of policy and practice. *Journal of Special Education Leadership*, *29*(2), 79–88. Retrieved from `https://www.learntechlib.org/p/192627`

Strubell, E., Ganesh, A., & McCallum, A. (2020). Energy and policy considerations for modern deep learning research. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 34, pp. 13693–13696).

Sun, Z., Yu, H., Song, X., Liu, R., Yang, Y., & Zhou, D. (2020). Mobilebert: A compact task-agnostic bert for resource-limited devices. *arXiv preprint arXiv:2004.02984*.

Suresh, H., & Guttag, J. (2021a, October). A framework for understanding sources of harm throughout the machine learning life cycle. In *Equity and access in algorithms, mechanisms, and optimization* (pp. 1–9). NY, USA: ACM. doi: 10.1145/3465416.3483305

Suresh, H., & Guttag, J. (2021b). Understanding potential sources of harm throughout the machine learning life cycle. *MIT Case Studies in Social and Ethical Responsibilities of Computing*(Summer 2021). doi: 10.21428/2c646de5.c16a07bb

Taiwo, E., Akinsola, A., Tella, E., Makinde, K., & Akinwande, M. (2023). *A review of the ethics of artificial intelligence and its applications in the United States.* (arXiv:2310.05751 [cs]) doi: 10.5121/ijci.2023.1206010

Tan, B., Armoush, N., Mazzullo, E., Bulut, O., & Gierl, M. (2024). A review of automatic item generation techniques leveraging large language models. *EdArXiv Preprints*. doi: 10.35542/osf.io/6d8tj

Tang, X. (2023). A latent hidden Markov model for process data. *Psychometrika*, *89*(1), 1–36. doi: 10.1007/s11336-023-09938-1

Trivedi, S. (2022). Improving students' retention using machine learning: Impacts and implications. *ScienceOpen Preprints*. doi: 10.14293/S2199-1006.1.SOR-.PPZMB0B.v2

Uto, M., Xie, Y., & Ueno, M. (2020). Neural automated essay scoring incorporating handcrafted features. In D. Scott, N. Bel, & C. Zong (Eds.), *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 6077–6088). Barcelona, Spain (Online): International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.535

Verdegem, P. (2021). *AI for everyone?: Critical perspectives.* University of Westminster Press.

Vincent-Lancrin, S., & Vlies, R. v. d. (2020). *Trustworthy artificial intelligence (AI) in education: Promises and challenges* (Tech. Rep.). Paris: OECD. doi: 10.1787/a6c90fa9-en

Vo, Y., Rickels, H., Welch, C., & Dunbar, S. (2023). Human scoring versus automated scoring for English learners in a statewide evidence-based writing assessment. *Assessing Writing*, *56*, 100719. doi: 10.1016/j.asw.2023.100719

Wan, S., & Keller, L. A. (2023). Using cumulative sum control chart to detect aberrant responses in educational assessments. *Practical Assessment, Research & Evaluation*, *28*. Retrieved 2024-05-22, from https://eric.ed.gov/?id=EJ1380532

Wang, C., Xu, G., Shang, Z., & Kuncel, N. (2018). Detecting aberrant behavior and item preknowledge: A comparison of mixture modeling method and residual method. *Journal of Educational and Behavioral Statistics*, *43*(4), 469–501. doi: 10.3102/1076998618767123

Ware, P. (2014). Feedback for adolescent writers in the english classroom. *Writing & Pedagogy*, *6*(2), 223–249. doi: 10.1558/wap.v6i2.223

Warschauer, M., & Grimes, D. (2008). Automated writing assessment in the classroom. *Pedagogies: An International Journal*, *3*(1), 22–36. doi: 10.1080/15544800701771580

Warschauer, M., & Ware, P. (2006). Automated writing evaluation: Defining the classroom research agenda. *Language Teaching Research*, *10*(2), 157–180. doi: 10.1191/1362168806lr190oa

Webber, K. L. (2019). The use and potential misuse of data in higher education: A compilation of examples. *IHE Research in Progress Series 2019-001*. Retrieved from https://ihe.uga.edu/sites/default/files/inline-files/Webber_2019001_paper.pdf

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of educational measurement*, *21*(4), 361–375. doi: 10.1111/j.1745-3984.1984.tb01040.x

Whitmer, J., Deng, E. Y., Blankenship, C., Beiting-Parrish, M., Zhang, T., & Bailey, P. (2023). *Results of NAEP reading item automated scoring data challenge (Fall 2021)*. OSF. Retrieved from https://osf.io/2hevq

Williamson, B., Macgilchrist, F., & Potter, J. (2023). Re-examining AI, automation and datafication in education. *Learning, Media and Technology*, *48*(1), 1–5. doi: 10.1080/17439884.2023.2167830

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, *31*(1), 2–13. Retrieved 2024-05-22, from https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1745-3992.2011.00223.x doi: 10.1111/j.1745-3992.2011.00223.x

Wilson, J., Ahrendt, C., Fudge, E. A., Raiche, A., Beard, G., & MacArthur, C. (2021). Elementary teachers' perceptions of automated feedback and automated scoring: Transforming the teaching and learning of writing using automated writing evaluation. *Computers & Education*, *168*, 104208. doi: 10.1016/j.compedu.2021.104208

Wilson, J., Chen, D., Sandbank, M. P., & Hebert, M. (2019). Generalizability of automated scores of writing quality in grades 3–5. *Journal of Educational Psychology*, *111*(4), 619–640. doi: 10.1037/edu0000311

Wilson, J., & Czik, A. (2016). Automated essay evaluation software in English language arts classrooms: Effects on teacher feedback, student motivation, and writing quality. *Computers & Education*, *100*, 94–109. doi: 10.1016/j.compedu.2016.05.004

Wilson, J., Palermo, C., & Wibowo, A. (2024). Elementary English learners' engagement with automated feedback. *Learning and Instruction*, *91*, 101890. doi: 10.1016/j.learninstruc.2024.101890

Wilson, J., & Roscoe, R. D. (2020). Automated writing evaluation and feedback: Multiple metrics of efficacy. *Journal of Educational Computing Research*, *58*(1), 87–125. doi: 10.1177/0735633119830764

Wilson, J., Zhang, F., Palermo, C., Cordero, T. C., Myers, M. C., Eacker, H., ... Coles, J. (2024). Predictors of middle school students' perceptions of automated writing evaluation. *Computers & Education*, *211*, 104985. doi: 10.1016/j.compedu.2023.104985

Wind, S. A., Wolfe, E. W., Engelhard Jr., G., Foltz, P., & Rosenstein, M. (2018). The influence of rater effects in training sets on the psychometric quality of automated scoring for writing assessments. *International Journal of Testing*, *18*(1), 27–49. doi: 10.1080/15305058.2017.1361426

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*(2), 163–183. doi: 10.1207/s15324818ame1802_2

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... others (2019). Huggingface's transformers: State-of-the-art natural language processing. *arXiv Preprint*. doi: 10.48550/arXiv.1910.03771

Wongvorachan, T., Lai, K. W., Bulut, O., Tsai, Y.-S., & Chen, G. (2022). Artificial intelligence: Transforming the future of feedback in education. *Journal of Applied Testing Technology*, 95–116. Retrieved from `http://jattjournal.net/index.php/atp/article/view/170387`

Yildirim-Erbasli, S. N., & Bulut, O. (2022). Designing predictive models for early prediction of students' test-taking engagement in computerized formative assessments. *Journal of Applied Testing Technology*. Retrieved 2024-05-22, from http://jattjournal.net/index.php/atp/article/view/167548

Yildirim-Erbasli, S. N., & Bulut, O. (2023). Conversation-based assessment: A novel approach to boosting test-taking effort in digital formative assessment. *Computers and Education: Artificial Intelligence*, *4*, 100135. doi: 10.1016/j.caeai.2023.100135

Yildirim-Erbasli, S. N., Bulut, O., Epp, C. D., & Cui, Y. (2023). Conversation-based assessments in education: Design, implementation, and cognitive walkthroughs for usability testing. *Journal of Educational Technology Systems*, *52*(1), 27-51. doi: 10.1177/00472395231178943

Yildirim-Erbasli, S. N., Gorgun, G., & Bulut, O. (2023). Enhancing self-regulated learning with artificial intelligence-powered learning analytics. In N. Kavaklı Ulutaş & D. Höl (Eds.), *Advances in early childhood and K-12 education* (pp. 57–83). IGI Global. doi: 10.4018/979-8-3693-0066-4.ch004

Yoder-Himes, D. R., Asif, A., Kinney, K., Brandt, T. J., Cecil, R. E., Himes, P. R., . . . Ross, E. (2022). Racial, skin tone, and sex disparities in automated proctoring software. *Frontiers in Education*, *7*(881449). doi: 10.3389/feduc.2022.881449

Zenisky, A., Hambleton, R. K., & Luecht, R. M. (2009). Multistage testing: Issues, designs, and research. In *Elements of adaptive testing* (pp. 355–372). Springer.

Zhang, M. (2013). Contrasting automated and human scoring of essays. *R & D Connections*, *21*(2), 1–11.

Zhang, M., Dorans, N., Li, C., & Rupp, A. (2017). Differential feature functioning in automated essay scoring. In H. Jiao & R. W. Lissitz (Eds.), *Test fairness in the new generation of large-scale assessment* (pp. 185–208). Information Age Publishing.

Zhang, M., Ruan, C., & Johnson, M. S. (2024). Explainable AI: Exploring subgroup differences in short-response scoring. *Paper presented at the annual meeting of the National Council for Measurement in Education, Philadelphia, PA.*

Zhang, N., Li, L., Chen, X., Deng, S., Bi, Z., Tan, C., . . . Chen, H. (2021). Differentiable prompt makes pre-trained language models better few-shot learners. *arXiv preprint*. doi: 10.48550/arXiv.2108.13161

Zhou, J., Chen, F., Berry, A., Reed, M., Zhang, S., & Savage, S. (2020). A survey on ethical principles of

AI and implementations. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)* (pp. 3010–3017). doi: 10.1109/SSCI47803.2020.9308437

Zhou, T., & Jiao, H. (2023). Exploration of the stacking ensemble machine learning algorithm for cheating detection in large-scale assessment. *Educational and Psychological Measurement*, *83*(4), 831–854. doi: 10.1177/00131644221117193

View publication stats