Medición y Evaluación del Aprendizaje Inclusivo. 2025; 2(2):16
doi: 10.63688/aprendizaje.v2.i2.16

REVISIÓN SISTEMÁTICA

Equity and bias in automated educational evaluation systems based on artificial intelligence

Equidad y sesgos en sistemas automatizados de evaluación educativa basados en inteligencia artificial

Pamela Navarro Escobar[1]

[1] Centro de Investigación, Desarrollo e Innovación JYMNIJA, Oruro, Bolivia

Submitted: 30-05-2025 | Revised: 07-06-2025 | Accepted: 19-08-2025 | Published: 29-08-2025

Corresponding Author: Pamela Navarro Escobar

ABSTRACT

Introduction: Equidad y sesgos en sistemas automatizados de evaluación educativa basados en inteligencia artificial is an emerging higher-education issue related to measurement quality, decision-making and learning experience. Objective: To systematically synthesize available evidence and identify patterns of effectiveness, methodological quality and implementation conditions. Method: A PRISMA-oriented systematic review was conducted on a consolidated academic corpus of 135 records; 7 studies met relevance and documentary sufficiency criteria. Results: Evidence was methodologically heterogeneous but converged on the need to align technology, constructs, feedback and educational decisions. Conclusions: Findings support evidence-informed adoption, institutional monitoring and continuous assessment of validity, equity and utility.

Keywords: higher education; systematic review; educational assessment; evidence; PRISMA; educational technology.

RESUMEN

Introducción: Equidad y sesgos en sistemas automatizados de evaluación educativa basados en inteligencia artificial representa un problema emergente para la educación superior por su relación con la calidad de la medición, la toma de decisiones y la experiencia de aprendizaje. Objetivo: sintetizar sistemáticamente la evidencia disponible y establecer patrones de efectividad, calidad metodológica y condiciones de implementación. Método: se aplicó una revisión sistemática con lógica PRISMA sobre un corpus académico consolidado de 135 registros, del cual se incluyeron 7 estudios que cumplieron criterios de pertinencia y suficiencia documental. Resultados: la evidencia muestra heterogeneidad de diseños y contextos, junto con convergencias en torno a la necesidad de alinear tecnología, constructo, retroalimentación y decisiones educativas. Conclusiones: los resultados respaldan una adopción basada en evidencia, seguimiento institucional y evaluación continua de validez, equidad y utilidad.

Palabras clave: educación superior; revisión sistemática; evaluación educativa; evidencia; PRISMA; tecnología educativa.

INTRODUCTION

University research on equity and biases in automated educational assessment systems based on artificial intelligence has become increasingly relevant because assessment decisions no longer depend exclusively on static instruments or in-person exchanges. Digitalization has expanded the amount of available evidence on performance, interaction, and learning, but it has also increased the need to distinguish between technically attractive innovations and practices capable of sustaining valid educational inferences. This review addresses this problem from an integrative perspective, focusing both on observed effects and on the methodological conditions that allow interpreting these effects without confusing technological availability with evaluative quality.

In this scenario, the literature on artificial intelligence applied to education warns that algorithmic bias can emerge at different stages of the data cycle and unequally affect student groups; fairness must therefore be examined as a property of the sociotechnical system rather than of the algorithm alone (8). In automated assessment, score validity also requires justifying the relevance of the construct, the representation of the task, the transparency of the model, and the consequences of using the results (9-11).

The conceptual core of the manuscript is organized around algorithmic justice. This approach avoids treating the literature as a simple collection of results and makes it possible to examine how each study defines the phenomenon, which population it observes, which instruments it uses, and what type of conclusion it considers legitimate. The heterogeneity of university contexts makes it necessary to pay attention to the relationship among design, implementation, and evidence, because the same strategy can produce benefits under certain conditions and modest results when teaching support, discipline, digital experience, or form of feedback changes.

This perspective also converges with recent frameworks of responsible AI in higher education, which articulate equity, accountability, transparency, and ethics as interdependent dimensions (15). Reviews of ethical risks in educational artificial intelligence indicate that non-discrimination, human oversight, responsible data management, and explainability are necessary conditions for justifiable institutional adoption (16,26).

This systematic review is pertinent because the field combines experimental studies, instrumental validations, observational analyses, technological developments, and previous syntheses. This diversity makes it difficult to draw solid conclusions from fragmentary reading. Instead of assuming that all studies address the same question, this article reconstructs patterns of convergence and divergence, identifies which results are repeated with greater consistency, and separates claims directly supported by the studies from those that still require additional validation. The purpose is to provide a useful reading for research, teaching, and university management.

Previous research on automated scoring confirms that technical accuracy alone does not resolve issues of validity, generalization, or comparability across tasks and populations. Reviews of essay scoring systems show an evolution from trait-based models toward neural architectures and language models, with ongoing challenges associated with coherence, construct representation, and transfer across contexts (12,13,30). In addition, explainability has been proposed as a requirement for enabling teachers, students, and institutional decision-makers to understand and question AI-assisted decisions (14).

Fairness, bias, and validity of automated decisions

The study (1) provided evidence published in 2026 through a methodological study of equity. It worked with multiple scoring models and demographic subgroups, which makes it possible to situate the scope of its inferences and assess their proximity to the university population of interest. In substantive terms, the work showed that equity varied by demographic group and level of competence, and the measures distinguished predictive bias from real differences. This antecedent is relevant for the algorithmic justice perspective because it shows that the interpretation of the effect depends on the correspondence between construct, measurement procedure, and application context. Its contribution is considered in this review as a piece of the comparative pattern, not as an isolated demonstration sufficient to generalize to all educational scenarios.

The study (2) provided evidence published in 2025 through empirical evaluation of algorithmic fairness. It worked with 38,722 text responses from PISA 2015, which makes it possible to situate the scope of its inferences and assess their proximity to the university population of interest. In substantive terms, the work found no discernible gender differences but did find a small bias by home language; it also identified item factors linked to inequity. This precedent is relevant to the algorithmic fairness perspective because it shows that the interpretation of the effect depends

on the correspondence between construct, measurement procedure, and application context. Its contribution is considered in this review as a piece of the comparative pattern, not as an isolated demonstration sufficient to generalize to all educational settings.

No. (3) provided evidence published in 2026 through empirical model comparison. The study used essays involving gender, ELL, and special education subgroups, which allows the scope of its inferences to be situated and its proximity to the target university population to be assessed. In substantive terms, the work showed that fine-tuned gPT-4o achieved the highest overall accuracy and fairness among the compared models. This antecedent is relevant for the algorithmic justice perspective because it shows that the interpretation of the effect depends on the correspondence between construct, measurement procedure, and application context. Its contribution is considered in this review as a piece of the comparative pattern, not as an isolated demonstration sufficient to generalize to all educational scenarios.

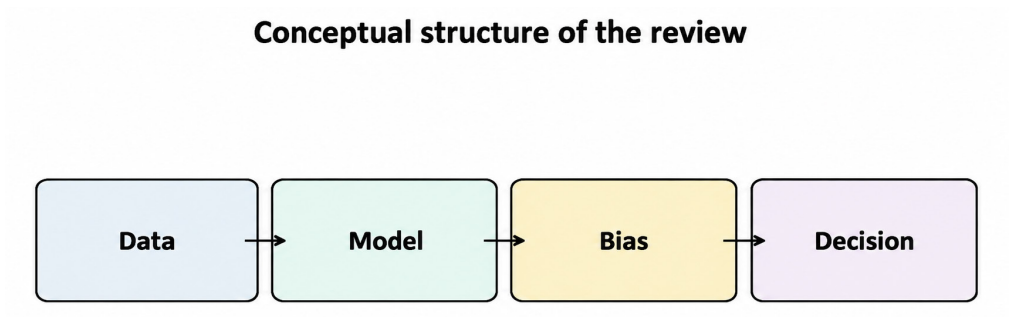
Ikkyu et al. (4) provided evidence published in 2025 through a methodological study with simulation and real data. The study used a real dataset, which allows the scope of its inferences to be situated and its proximity to the university population of interest to be assessed. In substantive terms, the study showed that the measure correctly identified characteristics responsible for bias in simulations and found consistent sets of variables in real data. This antecedent is relevant for the algorithmic fairness perspective because it shows that the interpretation of the effect depends on the correspondence between construct, measurement procedure, and application context. Its contribution is considered in this review as a piece of the comparative pattern, not as an isolated demonstration sufficient to generalize to all educational scenarios.

Matthew et al. (5) provided evidence published in 2022 through a methodological study with simulation and application. The study used simulations and a large-scale reading assessment item, which allows the scope of its inferences to be situated and its proximity to the university population of interest to be assessed. In substantive terms, the work showed how omission/proxies/confusion and optimization criteria can produce bias and how to evaluate it psychometrically. This antecedent is relevant to the algorithmic justice perspective because it shows that the interpretation of the effect depends on the correspondence between construct, measurement procedure, and application context. Its contribution is considered in this review as a piece of the comparative pattern, not as an isolated demonstration sufficient to generalize to all educational scenarios.

Hamida et al. (6) provided evidence published in 2026 through a technical accessibility evaluation. The study involved students with special educational needs in higher education, which allows the scope of its inferences to be situated and its proximity to the university population of interest to be assessed. In substantive terms, the work showed that digital assessment accessibility requires compatibility with assistive technologies, autonomy, and integrity under examination constraints. This antecedent is relevant to the algorithmic justice perspective because it shows that the interpretation of the effect depends on the correspondence between construct, measurement procedure, and application context. Its contribution is considered in this review as a piece of the comparative pattern, not as an isolated demonstration sufficient to generalize to all educational scenarios.

A. et al. (7) provided evidence published in 2022 through a methodological study with an illustrative application. The study used an illustrative application to automated video interviews, which allows the scope of its inferences to be situated and its proximity to the university population of interest to be assessed. In substantive terms, the work proposed a more flexible framework to incorporate stakeholder perspectives and equity in AI evaluations.

Figure 1. Conceptual structure of the review



Note. Own elaboration based on the analytical axes defined for the manuscript.

This antecedent is relevant for the algorithmic justice perspective because it shows that the interpretation of the effect depends on the correspondence between construct, measurement procedure, and application context. Its contribution is considered in this review as a piece of the comparative pattern, not as an isolated demonstration sufficient to generalize to all educational scenarios.

The figure organizes the article's reasoning as a conceptual sequence rather than as a universal template. The represented components correspond to the specific problem of this review and show the transition from the inputs or initial conditions to the educational interpretation. Its utility lies in making visible the relationships among constructs that, in individual studies, tend to appear fragmented across methodological and results sections. This representation guides the subsequent synthesis and facilitates distinguishing the elements that act as conditions, mechanisms, or results within the corpus.

Review questions

Question 1. What patterns of evidence allow understanding of dimension 1 of algorithmic justice in relation to fairness and biases in automated educational assessment systems based on artificial intelligence, and what methodological or contextual conditions modify its interpretation?

Question 2. What patterns of evidence allow understanding of dimension 2 of algorithmic justice in relation to fairness and biases in automated educational assessment systems based on artificial intelligence, and what methodological or contextual conditions modify its interpretation?

Question 3. What patterns of evidence allow understanding of dimension 3 of algorithmic justice in relation to fairness and biases in automated educational assessment systems based on artificial intelligence, and what methodological or contextual conditions modify its interpretation?

METHODS

PRISMA protocol

A systematic review was developed to synthesize the academic evidence relevant to the topic of the manuscript. The unit of analysis was the study or academic reference with a traceable source, identifiable methodology, and recoverable results. The consolidated academic working corpus consisted of 135 records previously gathered in the evidence matrices. A thematic and documentary selection was applied to this set to establish the final corpus. The procedure was organized according to the PRISMA logic of identification, screening, eligibility, and inclusion, adapted to the consolidated nature of the supplied sources.

The inclusion criteria required direct relevance to equity and biases in automated educational assessment systems based on artificial intelligence, correspondence with higher education or with instruments applicable to that level, sufficient description of the method, and availability of interpretable results. In total, 128 records were excluded because they had less relevance to the specific question, addressed peripheral populations or outcomes, or did not provide sufficient methodological evidence for the focused synthesis. The final corpus consisted of 7 studies. The selection was performed based on the information documented in the matrices, and no attribution of counts to any particular bibliographic database was made unless supported by the files.

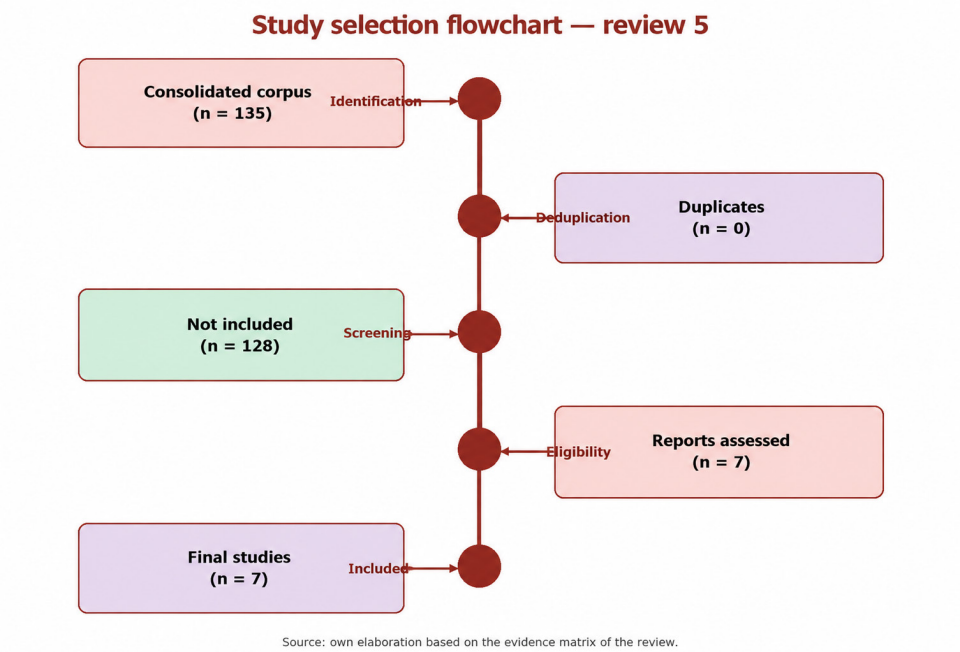
Data extraction was structured using a matrix that included authors, year, title, source, DOI or traceable link, provenance, study type, sample or corpus, summarized methodology, main results, level of evidence, and relevance. This organization enabled comparison of studies with heterogeneous designs without reducing them to a single metric. The synthesis combined descriptive characterization of the corpus with thematic analysis of the results, paying attention to convergences, contradictions, implementation conditions, and limits of generalization.

Evidence evaluation focused on the sufficiency of the retrieved information to answer the review question. Clarity of design, specification of the sample or corpus, correspondence between method and results, and traceability of the source were considered jointly. Meta-analysis was not performed when heterogeneity of outcomes, instruments, populations, or metrics precluded a responsible quantitative combination. In such cases, a comparative synthesis was prioritized that preserves differences among studies and avoids transforming non-equivalent results into a single estimator.

To strengthen transparency, the results were organized into characterization and synthesis tables, and the figures represent the selection flow and the temporal distribution of the corpus.

Each table was constructed from the fields of the supplied matrix. Subsequent interpretations were developed from observable patterns in the set, not from added assumptions. This criterion is especially important in educational reviews, where variation across disciplines, technologies, instruments, and institutional contexts can substantially modify the magnitude and meaning of the findings.

Figure 2. Flow of study selection according to PRISMA logic



Note. The flow represents 135 normalized academic records, without duplicates in the consolidated corpus; 128 records did not meet the thematic relevance threshold and 7 studies were included in the final synthesis.

The flow documents the screening and selection of the evidence. The 135 records in the consolidated corpus were normalized by DOI or title, so no duplicates were identified at this stage. The thematic screening excluded 128 records because of insufficient correspondence with the specific review question, whereas 7 studies retained methodological evidence and results sufficiently retrievable to be integrated into the synthesis. The sequence allows distinguishing documentary identification from the eligibility judgment and prevents the reduction of the corpus from being interpreted as an arbitrary loss of information.

RESULTS

Sources and forms of bias

The final corpus comprised 7 studies published during 2022–2026. The temporal distribution reveals an active and heterogeneous research field, with works corresponding to different stages of maturation of the problem. Some focus on demonstrating feasibility or association, whereas others advance toward validation, comparison of strategies, or analysis of implementation conditions. This temporal breadth makes it possible to observe not only favorable results but also shifts in how evaluation quality is conceptualized and in the type of evidence considered necessary to support educational decisions.

The methodological characterization confirms that equity and biases in automated educational assessment systems based on artificial intelligence do not constitute a homogeneous entity. The set includes designs with distinct purposes and variable levels of control over study conditions. This diversity broadens the scope of the review but also requires interpreting the findings according to the function of each design. Intervention studies report changes observed under defined conditions; instrumental studies provide evidence on measurement; observational analyses describe associations; and previous reviews allow identification of field trends without substituting for the evaluation of primary studies.

The samples and corpora described in the studies exhibit considerable variability. This variation is relevant because the transferability of the results depends on the size, composition, and context of the samples or corpora, as well as on the academic discipline and technological environment. Overall, the evidence suggests that the most interpretable results are those that specify the population, the evaluative task, and the mechanism by which the intervention or instrument should produce the expected effect. When any of these elements remains poorly defined, the utility of the finding for institutional decisions decreases.

The synthesis of results reveals a central pattern linked to algorithmic fairness: favorable effects or properties do not appear as automatic attributes of the technology or instrument, but as outcomes conditioned by design, implementation, data quality, and pedagogical alignment. The studies converge on the importance of articulating the tool with clear learning objectives, comprehensible evaluation criteria, and monitoring mechanisms. Divergences concentrate on the magnitude of effects, stability across contexts, and the capacity to sustain results when institutional conditions change.

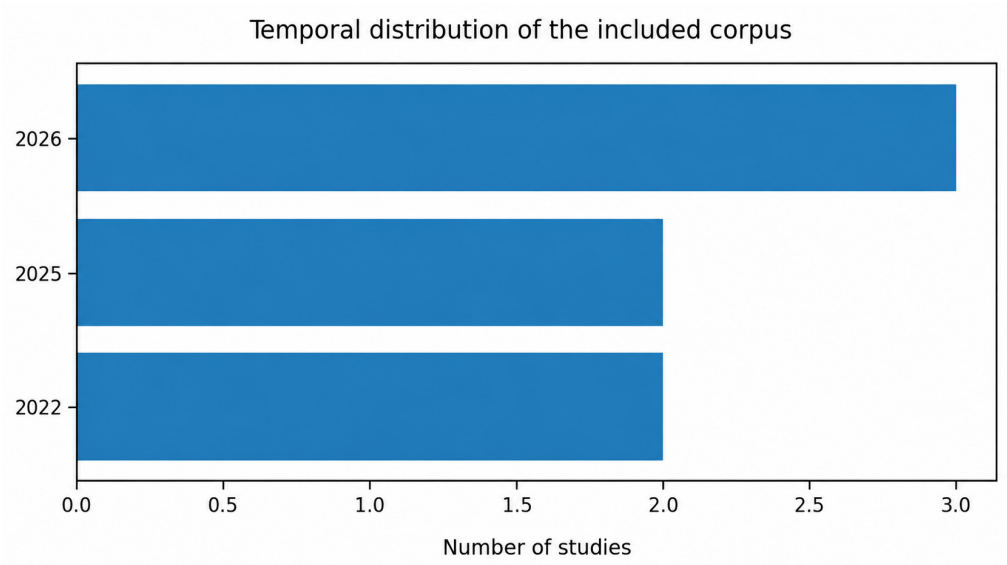
Table 1. Methodological characterization of the included corpus

Year	Design/methodology	Sample or corpus	Evidence
2026	Methodological study of equity	Multiple scoring models and demographic subgroups	High
2025	Empirical evaluation of algorithmic fairness	38,722 text responses from PISA 2015	High
2026	Empirical comparison of models	Trials with gender, ELL, and special education subgroups	High
2025	Methodological study with simulation and real data	Real dataset	High
2022	Methodological study with simulation and application	Simulations and a large-scale reading assessment item	High
2026	Technical accessibility evaluation	Students with special educational needs in higher education	High
2022	Methodological study/illustrative application	Illustrative application to automated video interviews	High

Note. Data extracted from the evidence matrix provided for this review.

The table shows the methodological diversity of the corpus and allows observing that the studies do not offer the same type of inference. The presence of designs with different objectives requires interpreting each result according to the question that the study can answer. This differentiation avoids directly comparing indicators that serve different functions and helps identify where real convergence exists. It also shows that the robustness of a conclusion depends on the combination of methodological clarity, sample description, and sufficiency of the evidence retrieved.

Figure 3. Temporal distribution of included studies



Note. Frequencies calculated from the year of publication recorded in the evidence matrix.

The temporal distribution makes it possible to situate the recent intensity of scientific production and to recognize whether the field is undergoing expansion, consolidation, or methodological transition. Rather than interpreting the number of publications as a measure of quality, the figure serves to contextualize the changes in questions, technologies, and instruments observed in the corpus. Reading it together with the methodological characterization allows us to identify whether the most recent works expand the evidence through more robust designs or whether exploratory approaches continue to predominate.

Mitigation and governance

Another cross-cutting pattern is the relevance of evidence quality. Studies with explicit procedures allow a better distinction among real improvement, favorable perception, and simple ease of use. In contrast, when the result relies mainly on self-report or short-term indicators, interpretation requires greater caution. The review shows that the practical usefulness of a strategy does not depend solely on its producing a statistical difference or a positive assessment, but on the measurement being consistent with the construct and on the intervention being able to be integrated sustainably into real educational processes.

In the subset of evidence corresponding to this axis, results were observed that, taken together, can be summarized as follows: Equity varied by demographic group and competence level; the measures distinguish predictive bias from real differences. No discernible gender differences were found, but a small bias by home language was observed; item factors linked to inequity were identified. Fine-tuned GPT-4o achieved the highest accuracy and overall fairness among the compared models. The comparative reading indicates that these findings should not be interpreted as equivalent, because they come from different designs and metrics. However, their thematic convergence allows the identification of recurring conditions and the formulation of an overall interpretation centered on the relationship between design quality, student response, and the usefulness of the information produced to guide educational decisions.

In the subset of evidence corresponding to this axis, results were observed that, considered jointly, can be summarized as follows: The measure correctly identified characteristics responsible for bias in simulations and found consistent sets of variables in real data. It showed how omission, proxies, confounding, and optimization criteria can produce bias and how to evaluate it psychometrically. It identifies that the accessibility of digital assessment requires compatibility with assistive technologies, autonomy, and integrity under examination constraints. The comparative reading indicates that these findings should not be interpreted as equivalent, because they come from different designs and metrics. However, their thematic convergence allows identifying recurrent conditions and formulating an overall interpretation focused on the relationship between design quality, student response, and the usefulness of the information produced for guiding educational decisions.

Table 2. Synthesis of relevant findings

Study	Recovered finding	Relevance
Assessing fairness in AI-assisted writing scoring: Developing fairness measures to detect	Fairness varied by demographic group and competency level; the measures distinguish predictive bias from real differences.	Included due to direct relevance.
Algorithmic Fairness in Automatic Short Answer Scoring	No discernible gender differences were found, but there was a small bias by home language; item factors linked to inequity were identified.	Included due to direct relevance.
Accuracy and fairness of generative AI in automated essay scoring: Comparing GPT-4o, featu	Fine-tuned GPT-4o achieved the highest precision and overall fairness among the compared models.	Included due to direct relevance.
Identifying Features Contributing to Differential Prediction Bias of Automated Scoring Sys	The measure correctly identified characteristics responsible for the bias in simulations and found consistent sets of variables in real data.	Included for its direct relevance.
Psychometric Methods to Evaluate Measurement and Algorithmic Bias in Automated Scoring	It showed how omission/proxies/confounding and optimization criteria can produce bias and how to evaluate it psychometrically.	Included due to direct relevance.
Accessible ICT-Enabled Examination Technologies for Students with Special Educational Need	It identifies that digital assessment accessibility requires compatibility with assistive technologies, autonomy, and integrity under examination constraints.	Included due to direct relevance.
Toward Argument-Based Fairness with an Application to AI-Enhanced Educational Assessments	It proposes a more flexible framework to incorporate stakeholder perspectives and equity in AI evaluations.	Included due to direct relevance.

Note. Textual synthesis based exclusively on the results recorded in the matrix.

In the subset of evidence corresponding to this axis, results were observed that, considered jointly, can be summarized as follows: It proposes a more flexible framework for incorporating stakeholder perspectives and equity into AI-based evaluations. The comparative reading indicates that these findings should not be interpreted as equivalent, because they come from different designs and metrics. However, their thematic convergence allows identifying recurrent conditions and formulating an overall interpretation focused on the relationship between design quality, student response, and the usefulness of the information produced for guiding educational decisions.

The synthesis shows that favorable evidence is distributed across different types of outcomes and that the relevance of each study depends on its proximity to the central construct. The findings are not interpreted as interchangeable: some describe performance, others measurement properties, student experience, accessibility, or model performance. The comparison allows recognizing convergences without losing the specificity of each result and constitutes the basis for discussing mechanisms and implementation conditions.

DISCUSSION

The synthesis allows us to argue that the problem of equity and bias in automated educational assessment systems based on artificial intelligence should be interpreted from a logic of algorithmic justice and not from a simplified opposition between effective and ineffective technology. The reviewed studies show that the performance of a strategy depends on the interaction between the assessment purpose, student characteristics, task design, measurement quality, and the institutional capacity to use the generated information. This conclusion shifts attention from the isolated tool to the assessment system in which it is embedded.

The corpus findings are consistent with external research showing that the choice of fairness metrics can modify the interpretation of the same predictive system. In analytical learning, there is no universally adequate metric: its selection depends on the context, the affected groups, and the type of decision to be supported (17,20). Therefore, the evaluation of algorithmic fairness must explicitly document which notion of fairness is adopted and what consequences prioritizing a metric over another would have.

An important consequence is that technological innovation does not eliminate the classic problems of validity, reliability, fairness, and utility. On the contrary, it can make them more visible by expanding the scale and speed of decisions. The reviewed evidence suggests that the best implementations are those that make criteria explicit, preserve opportunities for feedback, and maintain human review mechanisms when the consequences of a decision are significant. In this sense, digitalization should be understood as a transformation of the evaluative process and not as an automatic replacement of academic judgment.

Transparency and trust become especially relevant when algorithmic results guide interventions, grades, or academic support decisions. The literature on responsible learning analytics emphasizes that fairness, trust, transparency, and accountability must be incorporated from the design, not added only after detecting a disparity (18). Likewise, studies on historically underserved populations show that an algorithmic intervention may pursue equity goals while simultaneously introducing new ethical tensions if risk classification is not interpreted within its social context (19).

The observed heterogeneity also explains why it is not appropriate to reduce the field to a single effect size. Studies differ in constructs, duration, instruments, disciplines, samples, and metrics. This variation is not merely a limitation; it also provides information about the contexts in which a strategy works and about the conditions that must be considered before transferring it. Comparative synthesis allows patterns of consistency to be recognized without erasing differences that are essential for educational decision-making.

From a methodological standpoint, the corpus evidences a transition from studies focused on adoption and perception towards designs that seek to demonstrate measurement quality, impact on outcomes, and sustainability of the intervention. However, gaps persist related to longitudinal comparisons, replication across institutions, subgroup analyses, and documentation of unintended effects. The future agenda should prioritize designs that connect learning outcomes with implementation processes and that inform not only whether an intervention works, but for whom, under what conditions, and with what pedagogical or organizational costs.

Recent evidence also reinforces the need to evaluate the robustness of systems across subgroups. Studies on automated scoring have shown differences in accuracy and generalization depending on the dataset and the analyzed group, while research on student perception indicates

that acceptance of AI-based graders also depends on the perception of fairness of the procedure and the outcome (21,22). Consequently, technical fairness and perceived fairness should be considered related but not equivalent dimensions.

The main contribution of this review is to organize the evidence so that the results can be used for academic decisions without overextending their scope. The convergence across studies provides arguments for adopting promising practices, but methodological diversity requires preserving a logic of continuous evaluation. Institutions should treat each implementation as an intervention amenable to monitoring, with indicators defined before its deployment and mechanisms to review outcomes, biases, student experience, and coherence with curricular objectives.

Research on bias mitigation also shows that reducing a disparity does not guarantee simultaneously improving all fairness and performance metrics. Recent comparisons among reweighting, adversarial learning, and other procedures reveal trade-offs between predictive utility and equity among subgroups (23). This supports the advisability of reporting disaggregated results and conducting sensitivity analyses before using automated prediction in educational decisions with relevant consequences.

No. (1) offers a useful point of contrast for this interpretation, because its study documents that equity varied by demographic group and level of competence; the measures distinguish predictive bias from real differences. This finding acquires greater significance when compared with the rest of the corpus: it does not act as self-sufficient evidence, but rather as an observation that helps to specify mechanisms, limits, and conditions of application. The systematic reading thus makes it possible to avoid two frequent extremes—excessive generalization from a single study and the loss of contextual information when heterogeneous results are synthesized.

Reference (2) offers a useful point of contrast for this interpretation because its study found no discernible gender differences but did find a small bias by home language and identified item factors linked to inequity. This finding gains greater significance when compared with the rest of the corpus: it does not act as self-sufficient evidence, but rather as an observation that helps to clarify mechanisms, limits, and conditions of application. Systematic reading thus makes it possible to avoid 2 frequent extremes: excessive generalization from a single study and loss of contextual information when synthesizing heterogeneous results.

Reference (3) offers a useful point of contrast for this interpretation because its study documents that fine-tuned gPT-4o achieved the highest precision and overall fairness among the compared models. This finding gains greater significance when compared with the rest of the corpus: it does not act as self-sufficient evidence, but rather as an observation that helps to specify mechanisms, limits, and conditions of application. Systematic reading thus makes it possible to avoid 2 frequent extremes: excessive generalization from a single study and the loss of contextual information when synthesizing heterogeneous results.

Ikkyu et al. (4) offer a useful point of contrast for this interpretation because their study documents that the measure correctly identified characteristics responsible for bias in simulations and found consistent sets of variables in real data. This finding gains greater significance when compared with the rest of the corpus: it does not act as self-sufficient evidence, but rather as an observation that helps to clarify mechanisms, limits, and conditions of application. Systematic reading thus makes it possible to avoid 2 frequent extremes: excessive generalization from a single study and loss of contextual information when synthesizing heterogeneous results.

Matthew et al. (5) offer a useful point of contrast for this interpretation because their study documented and showed how omission/proxies/confounding and optimization criteria can produce bias and how to evaluate it psychometrically. This finding gains greater significance when compared with the rest of the corpus: it does not act as self-sufficient evidence, but rather as an observation that helps to specify mechanisms, limits, and conditions of application. Systematic reading thus makes it possible to avoid 2 frequent extremes: excessive generalization from a single study and the loss of contextual information when heterogeneous results are synthesized.

Hamida et al. (6) offer a useful point of contrast for this interpretation because their study documents and identifies that the accessibility of digital assessment requires compatibility with assistive technologies, autonomy, and integrity under exam restrictions. This finding gains greater significance when compared with the rest of the corpus: it does not act as self-sufficient evidence, but rather as an observation that helps to clarify mechanisms, limits, and conditions of application. Systematic reading thus makes it possible to avoid 2 frequent extremes: excessive generalization from a single study and the loss of contextual information when synthesizing heterogeneous results.

Safeguards for Implementation

The practical implications derived from the review should translate into gradual decisions. In addressing fairness and bias in automated educational evaluation systems based on artificial intelligence, responsible implementation requires defining in advance the construct to be measured, establishing success indicators, identifying potentially affected groups, and providing review mechanisms. The available evidence favors strategies that integrate innovation with teacher support and continuous evaluation of outcomes.

At the institutional level, adoption should be accompanied by data governance protocols, transparency criteria, and communication mechanisms with students and teachers. These conditions increase the interpretability of the results and reduce the risk of using indicators outside the context for which they were designed. The evaluation of a tool should consider both technical performance and pedagogical and distributional consequences.

Governance must also consider the biases present in educational content and data, as these can propagate to the models that use such content and data as input. Analytical learning has begun to incorporate specific methods to identify ethnic biases in educational materials, extending the assessment of equity beyond the final output of the algorithm (24). In higher education, the fairness of algorithmic evaluation is also related to student agency, data literacy, and the ability to understand how recommendations or classifications are generated (25).

For research, the next step is to better connect efficacy studies with implementation studies. It is necessary to know not only whether a strategy produces changes, but which components explain those changes, what resources it demands, and how it behaves in diverse populations. This orientation would allow evidence to be accumulated more coherently and reduce the fragmentation that currently characterizes part of the literature.

Contemporary reviews on AI in higher education agree that responsible assessment must integrate pedagogical and ethical dimensions throughout the instructional cycle, from data collection to the interpretation and use of results (27). In systems based on large language models, evaluation frameworks are likewise beginning to incorporate criteria of educational effectiveness, veracity, social alignment, and equity, while avoiding the reduction of the quality of the system to a single performance measure (31).

Evidence gaps

The main limitation of the corpus is its heterogeneity. Differences in designs, instruments, populations, and outcomes restrict the possibility of producing a common estimator and require that patterns be interpreted from a comparative perspective. This limitation does not invalidate the synthesis, but it conditions the scope of the conclusions and reinforces the need for replications with comparable protocols.

A concentration of studies is also observed in specific contexts and technologies, reducing certainty about transferability to institutions with different infrastructure, student profiles, or assessment cultures. Future research should document with greater precision the implementation conditions, participant characteristics, and decisions made during the deployment of the tools.

A particularly relevant gap corresponds to the representativeness of the training data. Recent evidence from automated writing scoring indicates that the demographic composition of the training set can modify bias patterns and that balanced sets can reduce, although not necessarily eliminate, disparities between groups (28). Similarly, comparisons between human raters and GPT models show that automated consistency can coexist with differences dependent on the evaluated criterion, reinforcing the need for construct validation and human supervision (29).

Finally, the review is based on the consolidated corpus documented in the provided matrices. The conclusions faithfully represent that set and do not extend to records that are not part of the available evidence. This delimitation preserves traceability and allows future updates to incorporate new studies without retrospectively altering the analytical basis used in the present manuscript.

CONCLUSIONS

The systematic review shows that equity and biases in automated educational assessment systems based on artificial intelligence constitute a field of study that, although heterogeneous in designs, populations, and metrics, has sufficient evidence to identify patterns. The central finding is that the quality of decisions depends less on the isolated presence of a technology or instrument than on its integration with learning objectives, explicit criteria, and monitoring procedures.

The algorithmic fairness perspective permits understanding of why similar interventions can produce different outcomes and why effectiveness evaluation must simultaneously consider results, conditions, and measurement quality.

In applied terms, higher education institutions should adopt a gradual implementation logic, with controlled testing, documentation of results, and periodic review of effects on different student groups. The evidence supports the use of digital strategies when there is pedagogical alignment and when the information generated is used to improve teaching and learning. It also shows the need to avoid automatic decisions based on single indicators, especially when academic consequences are relevant or when there are differences in access, technological experience, or support needs.

The research agenda must advance toward longitudinal studies, multicenter comparisons, and designs that document mechanisms and differential effects with greater precision. It is also necessary to improve the comparability of indicators and transparency in describing instruments, algorithms, and intervention procedures. These priorities will make it possible to transform a field characterized by rapid innovation into a more cumulative, replicable evidence base useful for university policies. The review provides a structure for this advance by separating consistent findings, implementation conditions, and gaps that still require research.

Expanded synthesis of the evidence

The evidence from a study in the corpus permits a deeper comparative interpretation. The work is characterized by the development of two equity measures; it compares ML, LLM, and group/proficiency differences and considers multiple scoring models and demographic subgroups. Among the retrieved results, equity varied by demographic group and proficiency level; the measures distinguish predictive bias from real differences. In the overall synthesis, this finding is used to contrast patterns and not to extend a conclusion beyond its design. Reading it together with other studies makes it possible to specify the relationship between measurement quality, implementation conditions, and educational utility, while also showing the need to distinguish between observed effects, user perceptions, and evidence on validity or performance. This differentiation strengthens the interpretation of the whole and prevents innovation from being evaluated solely on technological novelty.

The evidence from a study in the corpus permits a deeper comparative interpretation. The work combines semantic representations and classifiers; it compares disparities by gender and home language and considers 38,722 text responses from PISA 2015. Among the retrieved results, no discernible gender differences were found, but a small bias by home language was observed; the study also identified item factors linked to inequity. In the overall synthesis, this finding is used to contrast patterns and not to extend a conclusion beyond its design. Reading it together with other studies makes it possible to specify the relationship between measurement quality, implementation conditions, and educational utility, while also showing the need to distinguish between observed effects, user perceptions, and evidence on validity or performance. This differentiation strengthens the interpretation of the whole and prevents innovation from being evaluated solely on technological novelty.

The evidence from a study in the corpus allows for deeper comparative interpretation. The study compares GPT-4o with feature-based models and human raters, evaluating accuracy and fairness across subgroups; it examines essays from gender, ELL, and special education subgroups. The retrieved results show that adjusted GPT-4o achieved the highest overall accuracy and fairness among the compared models. In the global synthesis, this finding is used to contrast patterns and not to extend a conclusion beyond its design. Reading it together with other studies specifies the relationship between measurement quality, implementation conditions, and educational utility, while showing the need to distinguish between observed effects, user perceptions, and evidence on validity or performance. This differentiation strengthens the interpretation of the whole and prevents innovation from being evaluated solely by technological novelty.

The evidence from a study in the corpus allows for deeper comparative interpretation. The study proposes a bias contribution measure applicable to non-linear algorithms, tests it with neural networks and SVR, and considers a real dataset. The retrieved results show that the measure correctly identified features responsible for bias in simulations and found consistent sets of variables in real data. In the overall synthesis, this finding is used to contrast patterns and not to extend a conclusion beyond its design. Reading it together with other studies specifies the relationship between measurement quality, implementation conditions, and educational utility, while showing the need to distinguish between observed effects, user perceptions, and evidence on validity

or performance. This differentiation strengthens the interpretation of the whole and prevents innovation from being evaluated solely by technological novelty.

The evidence from a study in the corpus allows for deeper comparative interpretation. The study includes a review of bias mechanisms, two simple evaluation methods, simulations, and an empirical application with a large-scale reading evaluation item. The retrieved results show how omission/proxies/confusion and optimization criteria can produce bias and how to evaluate it psychometrically. In the overall synthesis, this finding is used to contrast patterns and not to extend a conclusion beyond its design. Reading it together with other studies specifies the relationship between measurement quality, implementation conditions, and educational utility, while showing the need to distinguish between observed effects, user perceptions, and evidence on validity or performance. This differentiation strengthens the interpretation of the whole and prevents innovation from being evaluated solely by technological novelty.

The evidence from a study in the corpus allows for deeper comparative interpretation. The study evaluates usability, reliability, quality of adjustments, and equity of digital examination platforms, focusing on students with special educational needs in higher education. The retrieved results show that digital assessment accessibility requires compatibility with assistive technologies, autonomy, and integrity under examination restrictions. In the overall synthesis, this finding is used to contrast patterns and not to extend a conclusion beyond its design. Reading it together with other studies specifies the relationship between measurement quality, implementation conditions, and educational utility, while showing the need to distinguish between observed effects, user perceptions, and evidence on validity or performance. This differentiation strengthens the interpretation of the whole and prevents innovation from being evaluated solely by technological novelty.

The evidence from a study in the corpus allows for deeper comparative interpretation. The study develops an equity argument complementary to validity and applies it to an AI-enhanced assessment, including an illustrative application to automated video interviews. The retrieved results show that the study proposes a more flexible framework to incorporate stakeholder perspectives and equity in AI-based assessments. In the overall synthesis, this finding is used to contrast patterns and not to extend a conclusion beyond its design. Reading it together with other studies specifies the relationship between measurement quality, implementation conditions, and educational utility, while showing the need to distinguish between observed effects, user perceptions, and evidence on validity or performance. This differentiation strengthens the interpretation of the whole and prevents innovation from being evaluated solely by technological novelty.

The evidence from a study in the corpus allows for deeper comparative interpretation. The study develops two equity measures, compares ML and LLM, examines differences across groups and proficiency levels, and considers multiple scoring models and demographic subgroups. The retrieved results show that equity varied by demographic group and proficiency level and that the measures distinguish predictive bias from real differences. In the global synthesis, this finding is used to contrast patterns rather than to extend a conclusion beyond its design. Reading it together with other studies specifies the relationship between measurement quality, implementation conditions, and educational utility, while showing the need to distinguish between observed effects, user perceptions, and evidence on validity or performance. This differentiation strengthens the interpretation of the whole and prevents innovation from being evaluated solely by technological novelty.

The evidence from a study in the corpus enables a deeper comparative interpretation. The study is characterized by its combination of semantic representations and classifiers; it compared disparities by gender and home language and considered 38,722 text responses from PISA 2015. Among the recovered results, no discernible gender differences were found, but a small home-language bias was observed, and item factors linked to inequity were identified. In the global synthesis, this finding is used to contrast patterns and not to extend a conclusion beyond its design. Reading the study together with other studies makes it possible to specify the relationship between measurement quality, implementation conditions, and educational utility, while also showing the need to distinguish between observed effects, user perceptions, and evidence on validity or performance. This differentiation strengthens the interpretation of the whole and prevents innovation from being evaluated solely by technological novelty.

The evidence from a study in the corpus enables a deeper comparative interpretation. The study is characterized by comparing gpt-4o with feature-based models and human raters; it evaluated accuracy and fairness across subgroups and considered essays for gender, ELL, and special education subgroups. Among the recovered results, fine-tuned gPT-4o achieved the highest overall accuracy and fairness among the compared models. In the global synthesis, this finding

is used to contrast patterns and not to extend a conclusion beyond its design. Reading the study together with other studies makes it possible to specify the relationship between measurement quality, implementation conditions, and educational utility, while also showing the need to distinguish between observed effects, user perceptions, and evidence on validity or performance. This differentiation strengthens the interpretation of the whole and prevents innovation from being evaluated solely by technological novelty.

The evidence from a study in the corpus enables a deeper comparative interpretation. The study is characterized by proposing a bias contribution measure applicable to nonlinear algorithms, testing with neural networks and SVR, and considering a real dataset. Among the retrieved results, the measure correctly identified features responsible for bias in simulations and found consistent sets of variables in real data. In the global synthesis, this finding is used to contrast patterns and not to extend a conclusion beyond its design. Reading the study together with other studies makes it possible to specify the relationship between measurement quality, implementation conditions, and educational utility, while also showing the need to distinguish between observed effects, user perceptions, and evidence on validity or performance. This differentiation strengthens the interpretation of the set and prevents innovation from being evaluated solely on technological novelty.

The evidence from a study in the corpus enables a deeper comparative interpretation. The study is characterized by a review of bias mechanisms, two simple evaluation methods, simulation, and empirical application, and it considers simulations and a large-scale reading assessment item. Among the results retrieved, the study showed how omission/proxies/confusion and optimization criteria can produce bias and how to evaluate it psychometrically. In the global synthesis, this finding is used to contrast patterns and not to extend a conclusion beyond its design. Reading the study together with other studies makes it possible to specify the relationship between measurement quality, implementation conditions, and educational utility, while also showing the need to distinguish between observed effects, user perceptions, and evidence on validity or performance. This differentiation strengthens the interpretation of the whole and prevents innovation from being evaluated solely by technological novelty.

The evidence from a study in the corpus enables a deeper comparative interpretation. The study is characterized by evaluating the usability, reliability, quality of adjustments, and equity of digital examination platforms, and it considers students with special educational needs in higher education. Among the retrieved results, the study identifies that the accessibility of digital assessment requires compatibility with assistive technologies, autonomy, and integrity under examination constraints. In the overall synthesis, this finding is used to contrast patterns and not to extend a conclusion beyond its design. Reading the study together with other studies makes it possible to specify the relationship between measurement quality, implementation conditions, and educational utility, while also showing the need to distinguish between observed effects, user perceptions, and evidence on validity or performance. This differentiation strengthens the interpretation of the whole and prevents innovation from being evaluated solely by technological novelty.

The evidence from a study in the corpus permits a deeper comparative interpretation. The study is characterized by developing an equity argument complementary to validity and applying it to an AI-enhanced assessment; it also considers an illustrative application to automated video interviews. Among the retrieved results, the study proposes a more flexible framework for incorporating stakeholder perspectives and equity in AI evaluations. In the overall synthesis, this finding is used to contrast patterns and not to extend a conclusion beyond its design. Reading the study together with other studies makes it possible to specify the relationship between measurement quality, implementation conditions, and educational utility, while also showing the need to distinguish between observed effects, user perceptions, and evidence on validity or performance. This differentiation strengthens the interpretation of the whole and prevents innovation from being evaluated solely on the basis of technological novelty.

REFERENCES

1. Chen G, Wu AD, Zhang C. Assessing fairness in AI-assisted writing scoring: Developing fairness measures to detect predictive bias in automated essay scoring. Assessing Writing. 2026;69:101066. https://doi.org/10.1016/j.asw.2026.101066
2. Andersen N, Mang J, Goldhammer F, et al. Algorithmic Fairness in Automatic Short Answer Scoring. International Journal of Artificial Intelligence in Education. 2025;35:3128–3165. https://doi.org/10.1007/s40593-025-00495-5
3. Huang Y, Palermo C, Wilson J. Accuracy and fairness of generative AI in automated essay scoring: Comparing GPT-4o, feature-based models, and human raters. Assessing Writing. 2026;69:101047. https://doi.org/10.1016/j.asw.2026.101047
4. Choi I, Johnson MS. Identifying Features Contributing to Differential Prediction Bias of Automated Scoring Systems. Journal of Educational Measurement. 2025;62(4):838–861. https://doi.org/10.1111/jedm.70015
5. Johnson MS, Liu X, McCaffrey DF. Psychometric Methods to Evaluate Measurement and Algorithmic Bias in Automated Scoring. Journal of Educational Measurement. 2022;59(3):338–361. https://doi.org/10.1111/jedm.12335
6. Al-Maamari H, Sidhu MS, Hussain SM, Al-Ghaili AM, Sidhu KK. Accessible ICT-Enabled Examination Technologies for Students with Special Educational Needs in Higher Education: A Technical Evaluation of Usability, Reliability, Accommodation Quality, and Assessment Fairness. International Journal of Special Education. 2026;41(17s):635–654.
7. Huggins-Manley AC, Booth BM, D'Mello SK. Toward Argument-Based Fairness with an Application to AI-Enhanced Educational Assessments. Journal of Educational Measurement. 2022;59(3):362–388. https://doi.org/10.1111/jedm.12334
8. Baker RS, Hawn A. Algorithmic Bias in Education. International Journal of Artificial Intelligence in Education. 2022;32(4):1052-1092. https://doi.org/10.1007/s40593-021-00285-9
9. Ferrara S, Qunbar S. Validity Arguments for AI-Based Automated Scores: Essay Scoring as an Illustration. Journal of Educational Measurement. 2022;59(3):288-313. https://doi.org/10.1111/jedm.12333
10. Ercikan K, McCaffrey DF. Optimizing Implementation of Artificial-Intelligence-Based Automated Scoring: An Evidence Centered Design Approach for Designing Assessments for AI-based Scoring. Journal of Educational Measurement. 2022;59(3):272-287. https://doi.org/10.1111/jedm.12332
11. Shermis MD. Anchoring Validity Evidence for Automated Essay Scoring. Journal of Educational Measurement. 2022;59(3):314-337. https://doi.org/10.1111/jedm.12336
12. Ramesh D, Sanampudi SK. An automated essay scoring systems: a systematic literature review. Artificial Intelligence Review. 2022;55(3):2495-2527. https://doi.org/10.1007/s10462-021-10068-2
13. Susanti MNI, Ramadhan A, Warnars HLHS. Automatic essay exam scoring system: a systematic literature review. Procedia Computer Science. 2023;216:531-538. https://doi.org/10.1016/j.procs.2022.12.166
14. Khosravi H, Shum SB, Chen G, Conati C, Tsai YS, Kay J. Explainable Artificial Intelligence in education. Computers and Education: Artificial Intelligence. 2022;3:100074. https://doi.org/10.1016/j.caeai.2022.100074
15. Memarian B, Doleck T. Fairness, Accountability, Transparency, and Ethics (FATE) in Artificial Intelligence (AI) and higher education: A systematic review. Computers and Education: Artificial Intelligence. 2023;5:100152. https://doi.org/10.1016/j.caeai.2023.100152
16. Zhu H, Sun Y, Yang J. Towards responsible artificial intelligence in education: a systematic review on identifying and mitigating ethical risks. Humanities and Social Sciences Communications. 2025;12:1111. https://doi.org/10.1057/s41599-025-05252-6
17. Deho OB, Zhan C, Li J, Liu J, Liu L, Le TD. How do the existing fairness metrics and unfairness mitigation algorithms contribute to ethical learning analytics? British Journal of Educational Technology. 2022;53(4):822-843. https://doi.org/10.1111/bjet.13217
18. Khalil M, Prinsloo P, Slade S. Fairness, Trust, Transparency, Equity, and Responsibility in Learning Analytics. Journal of Learning Analytics. 2023;10(1):1-7. https://doi.org/10.18608/jla.2023.7983
19. Froehlich L, Weydner-Volkmann S. Adaptive Interventions Reducing Social Identity Threat to Increase Equity in Higher Distance Education: A Use Case and Ethical Considerations on Algorithmic Fairness. Journal of Learning Analytics. 2024. https://doi.org/10.18608/jla.2024.8301
20. Cohausz L, Kappenberger J, Stuckenschmidt H. What Fairness Metrics Can Really Tell You: A Case Study in the Educational Domain. Proceedings of the 14th Learning Analytics and Knowledge Conference. 2024:792-799. https://doi.org/10.1145/3636555.3636873
21. Yang K, Raković M, Li Y, Guan Q, Gašević D, Chen G. Unveiling the Tapestry of Automated Essay Scoring: A Comprehensive Investigation of Accuracy, Fairness, and Generalizability. Proceedings of the AAAI Conference on Artificial Intelligence. 2024;38(20). https://doi.org/10.1609/aaai.v38i20.30254
22. Jones-Jang SM, Chung M, Choi J, Kim N, Lee S. Fairness perceptions of AI in grading systems: Examining how discontent with the status quo and outcome favorability reduce AI reluctance. Computers and Education: Artificial Intelligence. 2025;8:100419. https://doi.org/10.1016/j.caeai.2025.100419
23. Villegas-Ch W, Gutierrez R, Garcia-Ortiz J, Govea J. Mitigating algorithmic bias in educational prediction systems using reweighting and adversarial learning. Frontiers in Education. 2026;11:1841724. https://doi.org/10.3389/feduc.2026.1841724
24. Albuquerque J, Rienties B, Hlosta M, Holmes W. Learning Analytics to Uncover Ethnic Bias in Educational Texts: An Ensemble Learning Approach. Journal of Learning Analytics. 2026;13(1):143-162. https://doi.org/10.18608/jla.2026.8905

25. Lluch Molins L, Lindín Soriano C. The self-regulatory paradox of learning analytics: student expectations and the conditions for fair algorithmic assessment in higher education. Frontiers in Education. 2026;11:1913278. https://doi.org/10.3389/feduc.2026.1913278
26. Agarwal B, Urlings C, van Lankveld G, Klemke R. Identifying the ethical values and norms for artificial intelligence in education: A systematic literature review. International Journal of Artificial Intelligence in Education. 2026;36(1-2):100004. https://doi.org/10.1016/j.ijaied.2026.100004
27. Zhuang M, Long S, Martin F, Castellanos-Reyes D. The affordances of Artificial Intelligence (AI) and ethical considerations across the instruction cycle: A systematic review of AI in online higher education. The Internet and Higher Education. 2025;67:101039. https://doi.org/10.1016/j.iheduc.2025.101039
28. Holmes L, Morris W, Crossley S, Choi JS. Assessing fairness in finetuned scoring models with demographically restricted training data. Assessing Writing. 2026;68:101032. https://doi.org/10.1016/j.asw.2026.101032
29. Wu HN, Chu MN, Hsu JL. Comparing GPT and human raters in essay assessment: Variability, bias, and the potential of LLM-based scoring. Computers and Education Open. 2026;10:100341. https://doi.org/10.1016/j.caeo.2026.100341
30. Sun J, Song T, Weiming P, Song J. A survey of automated essay scoring: Challenges, advances, and future. Neurocomputing. 2025;650:130916. https://doi.org/10.1016/j.neucom.2025.130916
31. Lin P, Deng Q, Zhou Y. Towards responsible AI in education: A Delphi-AHP-based framework for evaluating educational large language models. Computers and Education: Artificial Intelligence. 2026;10:100534. https://doi.org/10.1016/j.caeai.2025.100534

FUNDING

None.

CONFLICT OF INTEREST

None exist.

AUTHORSHIP CONTRIBUTIONS

Conceptualization: Pamela Navarro Escobar
Data curation: Pamela Navarro Escobar
Formal analysis: Pamela Navarro Escobar
Research: Pamela Navarro Escobar
Methodology: Pamela Navarro Escobar
Supervision: Pamela Navarro Escobar
Validation: Pamela Navarro Escobar
Visualization: Pamela Navarro Escobar
Writing – original draft: Pamela Navarro Escobar
Writing – review and editing: Pamela Navarro Escobar