Multidimensional Research Insights
Volume 2, Issue 1
ISSN: 3067-8129

Equity and Bias in AI-Based Educational Assessments: Impacts on SEND Learners, Gifted Students, and Gender Representation in UAE Schools

[1]*Athira B Nair
[1]Head of Assessment. And pursuing PHD in Paris American International University, France.

Abstract

Background: The rapid integration of AI (AI) into educational assessment systems raised concerns regarding fairness, equity, and algorithmic bias, particularly for vulnerable student populations.

Purpose: This study investigated equity and bias in AI-based educational assessments within the United Arab Emirates (UAE), focusing on SEND learners, gifted students, and gender-diverse groups.

Methods: A convergent parallel mixed-methods design employed purposive sampling from 15 UAE schools. Participants included 400 students (ages 11-18 years, M=14.3, SD=2.31; 192 males, 185 females, 20 non-binary), 82 teachers (56 females, 24 males), and 28 school leaders. Quantitative data were analyzed using SPSS (t-tests, ANOVA, regression, Cohen's d). Qualitative data underwent thematic analysis (Cohen's kappa=.84).

Results: SEND students experienced severe disadvantage across equity dimensions (d=0.76-1.12). Gifted students rated pedagogical value significantly lower (d=0.61). Female students perceived AI assessment as less fair than males (d=0.45). Teacher trust correlated exceptionally with transparency (r=.970, R^2=.941). Hybrid human-AI models were universally preferred (M=4.25).

Conclusion: AI assessments introduced systematic biases disadvantaging marginalized learners. Achieving equity required fairness auditing, transparent algorithms, inclusive datasets, robust governance, and preserved human judgment.

Keywords: AI, Educational Assessment, Algorithmic Bias, SEND Learners, Gender Equity, UAE Education

1 Introduction

The fast evolution of AI (AI) changed many areas of the world, and education was not an exception (Kou et al., 2024; Mateos & Bellogín, 2024). Educational testing, which was traditionally based on human-based assessment practices, became aligned at the border of technological innovation and pedagogical tradition. There was a growing concern about equity, fairness, transparency, and protection of educational values which focused on human dignity and individual differences as AI systems were more often involved in the evaluation and measurement of students and in the provision of academic support (Williams, 2024). Educational testing formed the basis of structural processes in academic institutions, which were involved in curriculum development, policy making, and student academic development (Daniel & Onwuegbuzie, 2007).

Traditional measures of assessment, such as standardized tests and portfolio-based assessments, were designed to test cognitive skills, understanding, and problem-solving skills. Nevertheless, they tended to be ineffective in supporting the plurality of learning profiles that were typical of modern student bodies, such as the learners with SEND, gifted students who needed an alternative level of challenge, and gender-diverse learners (Signorella, 2020). The opportunities offered by AI-based technologies were the ability to overcome these constraints by automation, personalization, and scalability. Learning analytics and machine learning algorithms allowed providing feedback in real-time, creating adaptive evaluation tracks, and evaluation mechanisms that responded to the student performance dynamically (Mateos and Bellogin, 2024). Market prognoses indicated that the global AI in education market would amount to about USD 32.27 billion by 2030, serving as a sign of rising usage of AI-based tools like automated models of scoring essays, intelligent tutoring systems, and adaptive testing platforms (Kou et al., 2024). Though such advantages could be observed, AI as a tool in assessments in education presented significant ethical, technical, and social issues. The most pressing issue was that of algorithmic bias the systemic distortions inherent

Athira B Nair
Head of Assessment. And pursuing PHD in Paris American International University, France.
Email: athirabnair99@gmail.com

Received: 16-Feb-2026 Revised: 19-Mar-2026 Accepted: 31-Mar-2026

in the AI systems that potentially replicated or even enlarged the inequities that already existed (Williams, 2024). Training AI models using past biased or non-representative data exposed them to the risk of reinstating these patterns of discrimination that disproportionately impacted marginalized groups of learners (Neto et al., 2024). The United Arab Emirates (UAE) provided a unique background on which to analyze these dynamics. The UAE was a pioneer in the use of AI in the region, and education became a strategic direction, which was reflected in its National AI Strategy 2031. The overlap of AI-based assessment solutions and the inclusive education objectives in the UAE prompted the opportunity to customize educational practices, as well as the problem of equity, governance, and ethical application.

This study addressed five research questions: (RQ1) To what extent did AI-based educational assessment systems used in UAE schools demonstrate equity or bias toward SEND learners, gifted students, and different gender groups? (RQ2) How effectively did AI-based assessments accommodate diverse learning needs, including accessibility requirements, differentiation demands, and adaptive challenge provision? (RQ3) What forms of algorithmic bias, if any, were perceived or evidenced in AI-based assessment outcomes related to gender representation and intersectional identities? (RQ4) How did teachers, students, and parents perceive the fairness, transparency, and trustworthiness of AI-based assessments deployed in UAE schools? (RQ5) What governance, pedagogical, and ethical frameworks were needed to ensure equitable AI-based assessment practices in UAE educational contexts?

2 Literature Review

The use of AI in educational assessment was changing current educational assessment practices at a rapid pace, presenting both opportunities to achieve unprecedented levels of personalization and threatening challenges of fairness and equity. This literature review provided a detailed analysis of published literature applicable to equity and bias in AI-based educational assessment, with particular focus on SEND learners, gifted students, and gender representations in UAE schools.

2.1 Theoretical Framework

Three theoretical perspectives formed the basis of this research: algorithmic fairness frameworks, UDL, and socio-technical governance models. The combination of these views created a rich conceptual base to study the ways AI-based assessment systems could either facilitate or hinder educational equity (Holstein et al., 2019; Layachi & Pitchford, 2024) UNESCO, 2023).

2.1.1 Algorithmic Fairness Frameworks

Algorithmic fairness frameworks conceptualize equity as minimizing bias across the measurement, modeling, and action stages of AI development (Boateng & Boateng, 2025; Zhang, 2025). Bias may arise from unrepresentative training data, flawed assumptions, or feedback loops that reinforce historical inequalities (Owan et al., 2023). Kizilcec and Lee (2020) identified key decision points where learner characteristics, algorithm design, and system implementation shape outcomes (Kizilcec & Lee, 2020). Contemporary research extends fairness beyond equal treatment toward equity-focused evaluation, emphasizing disaggregated analysis (Yang et al., 2024). A systematic review further documented persistent disadvantages for marginalized groups across educational AI applications (Baker, 2022), underscoring the need for multidimensional fairness assessment.

2.1.2 Universal Design for Learning

Universal Design for Learning (UDL) complements algorithmic fairness by embedding inclusion through multiple means of engagement, representation, and expression (Layachi & Pitchford, 2024; Majumdar, 2025). Originating from universal design in architecture, UDL treats learner variability as normative and promotes proactive, rather than reactive, accommodation. AI technologies theoretically support UDL through adaptive feedback, customizable formats, and multimodal learning pathways (Saputra et al., 2024; Yakkala, 2024). However, empirical implementation remains limited. Ethical integration must address risks including bias, privacy, and academic misuse (Mallary, 2025). Effective AI assessment design therefore requires flexible pacing, scaffolded supports, and differentiated difficulty to meet SEND and gifted learners' diverse cognitive, sensory, and linguistic needs.

2.1.3 Socio-Technical Governance Models

Socio-technical governance models extend fairness beyond technical design to institutional accountability, oversight, and ethical deployment (Holstein et al., 2019; Unesco, 2021). These frameworks recognize AI systems as embedded within complex social and organizational contexts, where technical safeguards alone are insufficient.

Empirical evidence shows that sustained fairness depends on continuous monitoring, stakeholder engagement, and iterative refinement (Holstein et al., 2019). Governance principles emphasize transparency, documentation of system capabilities and limits, mechanisms to contest automated decisions, and preservation of human control in high-stakes settings (Unesco, 2023). In the UAE, such governance is critical to align rapid AI expansion with inclusive education policies (El Naggar et al., 2024).

2.2 Conceptual Framework

Figure 3 represented the conceptual framework of this study and showed how the combination of algorithmic fairness mechanisms, UDL-informed assessment design, and socio-technical governance structures determined the equitable AI-based assessment results (Holstein et al., 2019; Layachi & Pitchford, 2024; Unesco, 2023). The framework illustrated that equitable AI assessment was recognized as a multidimensional phenomenon that could not be confined to any single theoretical perspective; technical, pedagogical and institutional considerations were all necessary to achieve equity. The three pillars were not separate areas but interacted dynamically, with developments in one area having implications for the possibilities and constraints in the others. The framework was grounded in empirical data from the UAE context, with theoretical constructs not treated as abstract categories but as means of interpreting concrete data.

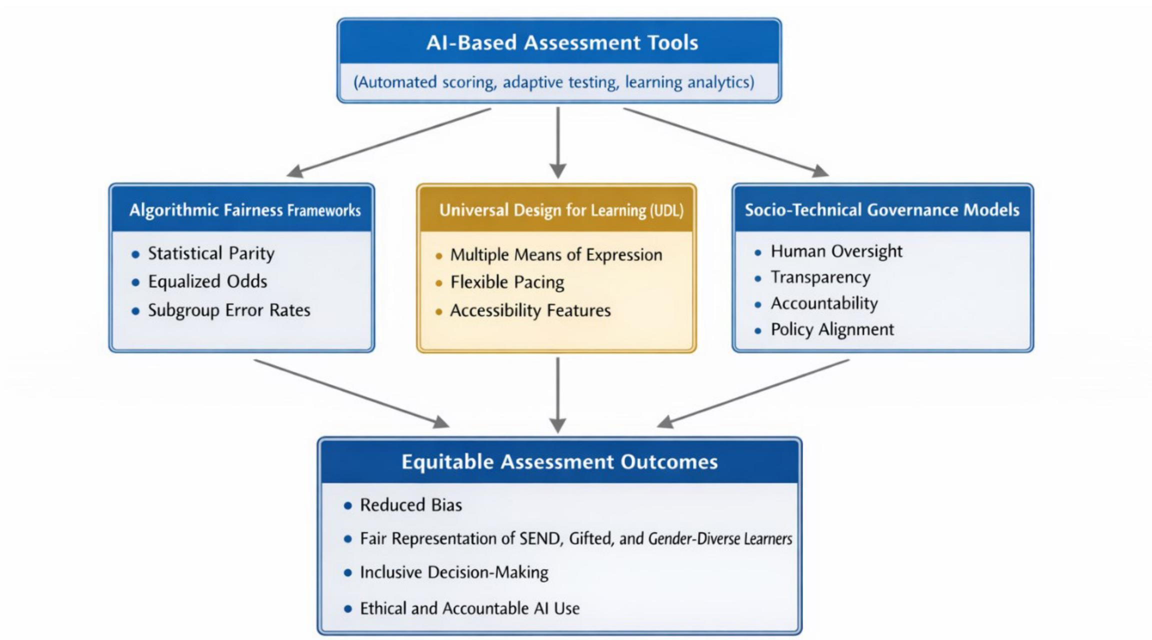

Figure 1: Conceptual Framework for Equitable AI-Based Educational Assessment

Figure 1 Conceptual framework diagram showing three pillars: (1) Algorithmic Fairness Theory, (2) Universal Design for Learning, and (3) Socio-Technical Governance Models, all leading to Equitable AI-Based Assessment, grounded in UAE Educational Context including AI platform analytics, stakeholder perceptions, and institutional practices.

2.3 Thematic Review of Literature

2.3.1 Fairness and Validity in AI-Based Assessment

Automated Essay Scoring (AES) represents one of the most established AI applications in educational assessment. Early systems such as e-rater demonstrated reliability comparable to human raters (Attali & Burstein, 2006), using linguistic and structural text features. However, subsequent research identified demographic fairness concerns. Different AES architectures exhibited distinct bias patterns across gender, race, and socioeconomic groups, with no universally equitable model (Litman & Madnani, 2021). Large-scale evaluation showed prompt-specific models improved accuracy but increased socioeconomic bias, while simpler models sometimes balanced fairness better (Yang et al., 2024). Overreliance on human benchmarks may perpetuate inequities and SEND and gifted learners remain underexamined (Williams, 2024).

Multidimensional Research Insights
Volume 2, Issue 1
ISSN: 3067-8129

2.3.2 Inclusion, SEND, and Gifted Learners

AI assessment within inclusive education for SEND and gifted learners remains insufficiently studied. Although AI enables adaptive difficulty, flexible formats, and personalized support, empirical evidence on equitable outcomes is limited (Unesco, 2021). Inclusive policy frameworks require full participation of all learners in mainstream environments (Khda, 2017) KHDA, necessitating cognitive, sensory, and linguistic accessibility. Poorly designed systems risk reinforcing barriers for SEND students. Conversely, gifted learners often encounter ceiling effects in adaptive assessments, limiting challenge and accurate ability representation (Baker et al., 2016). These gaps highlight the importance of differentiated, evidence-based AI design to ensure assessments support diverse learning needs without introducing new inequities.

2.3.3 Gender and Intersectional Representation

Gender bias in AI systems is widely documented across language and vision technologies. Word embeddings encode stereotypical associations that link male identities with careers and female identities with domestic roles, potentially shaping biased educational tools (Caliskan et al., 2017). Automated assessment may therefore privilege certain communication styles. Empirical findings are mixed, with some models showing neutrality while others reveal disparities related to language background (Ijaied, 2025). Gender-diverse learners face heightened risks due to binary demographic classifications that obscure representation. Intersectional effects involving SEND status, giftedness, and socioeconomic background remain largely unexplored, despite evidence of differential outcomes in adaptive systems (Holstein & Doroudi, 2021).

2.3.4 Policy, Governance, and Ethical AI

International frameworks increasingly stress ethical governance for AI in education. UNESCO (2023) emphasizes transparency, accountability, human agency, and mechanisms to challenge automated decisions, while preserving human oversight in high-stakes contexts. The OECD (2023) advocates representative training data, disaggregated fairness audits, algorithmic disclosure, and accessible appeals processes. In the UAE, the National AI Strategy 2031 promotes large-scale AI integration across education (UAE Cabinet, 2019), accelerating innovation but increasing equity risks. Concurrently, the Dubai Inclusive Education Policy Framework supports shared environments and individualized support (Khda, 2017). Effective governance must therefore align technological adoption with inclusion, ensuring AI enhances rather than compromises fairness.

2.4 Hypothesis Development

Based on theoretical frameworks, prior research on algorithmic bias, Universal Design for Learning (UDL), and socio-technical governance, this study proposes five hypotheses examining equity in AI-based educational assessments.

H1 predicts that AI systems show lower accuracy and fewer accommodations for SEND students due to underrepresentation in training data.

H2 suggests gifted learners experience reduced equity and engagement because limited adaptive challenge creates ceiling effects.

H3 anticipates gender bias in scoring and feedback, reflecting historical and imbalanced training datasets, particularly disadvantaging gender-diverse students.

H4 proposes that teachers' trust in AI assessments increases with system transparency and explainability, while opaque algorithms reduce confidence regardless of performance.

H5 expects stakeholders to perceive hybrid human-AI assessment models as more equitable than fully automated systems, as human oversight better addresses diverse learner needs and contextual judgment.

Despite the growing body of research on AI in education, several critical gaps remained that this study aimed to address. First, existing studies had underrepresented K-12 educational contexts and non-Western settings, with most research focused on Western institutions of higher learning (Holmes et al., 2022; Williamson & Eynon, 2020). The UAE provided a unique case due to its National AI Strategy 2031 and commitment to inclusive education, yet remained empirically underexplored. Second, SEND learners, gifted students, and gender-diverse populations had received minimal focused attention, with little intersectional analysis of how ability, gender, and other identities interacted (Kizilcec & Lee, 2022). Third, methodologically, few studies adopted mixed-method research designs combining algorithmic audits with stakeholder perspectives (Mehrabi et al., 2021). Fourth, practical frameworks to assist schools in auditing AI systems, tracking fairness, and adjusting tools to local contexts were largely absent (Floridi et al., 2018; Unesco, 2023). These gaps collectively justified the current investigation's focus on SEND learners, gifted students, and gender diversity within UAE schools, employing a mixed-methods design to generate both statistical evidence and contextual understanding.

3 Methods

3.1 Research Design

The research employed a convergent parallel mixed-methods design to explore equity and bias in AI-based educational assessment in UAE educational settings. Quantitative and qualitative data were collected simultaneously and analyzed together in the analysis and interpretation stages. The quantitative component enabled systematic analysis of assessment results, performance trends, and potential biases across subgroups of learners, including students with SEND, gifted students, and students with different gender identities. The qualitative element provided contextual richness by examining stakeholders' views, life experiences, and decision-making documents from institutions regarding the adoption of AI assessment.

3.2 Participants

The target group included students, teachers, and school leaders from 15 UAE-based schools that actively used AI-assisted assessment tools. A multi-stage purposive sampling strategy was used to ensure contextual relevance and demographic coverage. Schools were selected based on active use of AI-based assessment devices for at least one academic year and evidence of inclusive education practices through published inclusion policies and SEND-support arrangements. The sampling was done across schools following various curricular systems, including American, British, and International Baccalaureate schools, with diverse socio-economic groups.

The final sample consisted of 400 students in grades 6 to 12. The sample comprised 202 students with SEND (50.5%), 102 gifted-and-talented students (25.5%), and the remainder without special educational designations. Male students comprised 48% (n=192), female students 46.2% (n=185), non-binary students 5% (n=20), and 0.8% (n=3) preferred not to disclose. Students ranged from 11 to 18 years (M=14.3, SD=2.31). The teacher sample comprised 82 educators across the same 15 participating schools. Female teachers constituted 68.3% (n=56), male teachers 29.3% (n=24), and 2.4% (n=2) identified as other gender identities. Nearly two-thirds of teachers (64.6%, n=53) reported extensive or moderate experience working with SEND students. The school leadership sample comprised 28 administrators and technology coordinators: school principals (46.4%, n=13), assistant principals (25%, n=7), curriculum coordinators (17.9%, n=5), and technology specialists (10.7%, n=3).

3.2.1 Data Collection

Quantitative data were collected in participating schools across the UAE using surveys, AI platform analytics, and quasi-experimental comparisons. Structured electronic questionnaires were administered to students, teachers, and administrators to assess perceived fairness, accuracy, accessibility, trust, and pedagogical value using five-point Likert scales. Surveys were available in English and Arabic, with translation and back-translation ensuring linguistic equivalence. Anonymized AI platform analytics provided objective performance indicators, including scoring distributions, feedback patterns, adaptive difficulty, and completion rates. Subgroup analyses examined differential outcomes. Parallel assessments were graded independently by AI systems and blinded human raters to enable direct comparison of reliability, validity, and equity. Qualitative data were gathered through semi-structured interviews conducted in UAE schools to explore stakeholder experiences with AI-based assessment. Nineteen primary interviews were completed with students, teachers, and school leaders, followed by seven additional follow-up interviews. Sessions lasted 45–75 minutes (M = 58 minutes) and were conducted in participants' preferred language (English or Arabic). Interviews were audio-recorded with informed consent, transcribed verbatim, and translated when necessary. Protocols examined perceptions of fairness, accessibility, transparency, trust, motivational impacts, and improvement needs. This approach enabled in-depth, contextualized understanding of how AI assessment practices affected diverse educational stakeholders.

3.2.2 Data Analysis

Quantitative analyses were conducted using SPSS Statistics Version 29 to ensure systematic and reproducible procedures. Data screening addressed missing values, outliers, and distributional assumptions; cases exceeding 20% missing data were removed, and remaining gaps were treated using multiple imputation. Descriptive statistics summarized demographics, response patterns, and central tendencies. Inferential analyses included independent-samples t-tests, ANOVA with Bonferroni-adjusted post-hoc comparisons, and multiple regression examining associations between demographic and contextual predictors (e.g., SEND status, gender, prior achievement, AI exposure) and outcomes (e.g., fairness perceptions, assessment scores, trust). Chi-square tests assessed

Multidimensional Research Insights
Volume 2, Issue 1
ISSN: 3067-8129

categorical relationships. Effect sizes (Cohen's d, eta-squared, Cramér's V) complemented significance testing to evaluate practical relevance.

3.2.3 Ethical Considerations

Ethical approval was secured from the institutional review board prior to data collection, and the study adhered to established ethical research principles throughout. All participants received clear information about the study's purpose, procedures, risks, and data handling practices. Written informed consent was obtained from adult participants, while parental consent and student assent were secured for participants under 18. Consent materials were provided in English and Arabic and adapted for SEND participants using accessible formats. Participant privacy and confidentiality were ensured through pseudonymization, aggregated reporting, and secure storage of encrypted digital data and locked physical records. SEND participants received appropriate accommodation aligned with their educational plans. Participation was voluntary, with the right to withdraw at any stage without penalty. All procedures complied with UAE data protection regulations and international standards for research involving minors.

3.3 Results

3.3.1 Scale Reliability

Before hypothesis testing, internal consistency reliability for all multi-item scales was assessed using Cronbach's alpha. Table 1 presented reliability statistics for student perception scales. All scales showed exceptional internal consistency, with Cronbach's alpha coefficients exceeding .87. The majority of scales achieved $\alpha > .96$, indicating excellent reliability and supporting scale construction. These high reliabilities suggested that items within each scale measured coherent, unidimensional constructs. The Perceived Fairness scale demonstrated $\alpha = 0.962$ across 8 items. The Assessment Accuracy scale showed $\alpha = 0.963$ across 7 items. The Accessibility & Inclusiveness scale exhibited $\alpha = 0.966$ across 9 items, representing the highest reliability. The Trust & Transparency scale demonstrated $\alpha = 0.963$ across 8 items. The Pedagogical Value scale showed $\alpha = 0.965$ across 6 items. The AI vs Human Preference scale, while slightly lower, still demonstrated good reliability at $\alpha = 0.872$ across 5 items.

Table 1: Internal Consistency Reliability for Student Scales

Scale	Items	Cronbach's α	Interpretation
Perceived Fairness	8	0.962	Excellent
Assessment Accuracy	7	0.963	Excellent
Accessibility & Inclusiveness	9	0.966	Excellent
Trust & Transparency	8	0.963	Excellent
Pedagogical Value	6	0.965	Excellent
AI vs Human Preference	5	0.872	Good

3.3.2 AI Bias Toward SEND Students

The first hypothesis predicted that AI assessment systems showed statistically significant bias in outcomes for SEND students compared with neurotypical peers. Independent samples t-tests were conducted to compare SEND and non-SEND students across four equity-related perception scales. Results were presented in Table 2 and visualized in Figure 2. Results provided overwhelming support for the hypothesis. SEND students scored significantly lower than their non-SEND peers across all four perception dimensions, with all differences reaching statistical significance at $p < .001$. Effect sizes ranged from large ($d = 0.76$ for Perceived Fairness) to very large ($d = 1.12$ for Accessibility & Inclusiveness), indicating both statistical significance and substantive practical importance.

The most pronounced disparity emerged in Accessibility & Inclusiveness ($d = 1.12$), with SEND students rating AI systems nearly 0.8 points lower on the 5-point scale than non-SEND students (SEND $M = 2.70$, $SD = 0.71$; Non-SEND $M = 3.48$, $SD = 0.69$; $t(398) = -11.18$, $p < .001$). This very large effect size suggested that AI assessment tools fundamentally failed to provide equitable access for students with learning disabilities. The mean of 2.70 for SEND students fell well below the neutral midpoint of 3.0, indicating that SEND students perceived AI assessment as actively inaccessible and non-inclusive. For Perceived Fairness, SEND students averaged 2.58 (SD =

0.68) compared to non-SEND students' 3.11 (SD = 0.72), representing a statistically significant difference with a large effect (t(398) = -7.62, p < .001, d = 0.76). Similarly, Assessment Accuracy showed SEND students rating the systems at 2.72 (SD = 0.72) versus non-SEND students at 3.29 (SD = 0.72), yielding t(398) = -7.89, p < .001, d = 0.79. Trust & Transparency demonstrated SEND students scoring 2.46 (SD = 0.72) compared to non-SEND students' 3.02 (SD = 0.73), with t(398) = -7.73, p < .001, d = 0.77. Figure 4 provided visual confirmation of this systematic pattern, with SEND students consistently and substantially lower than non-SEND students across all four dimensions.

Table 2: Independent t-test Results - SEND vs Non-SEND Students

Scale	SEND M (SD)	Non-SEND M (SD)	t	df	p	Cohen's d
Perceived Fairness	2.58 (0.68)	3.11 (0.72)	-7.62	398	<.001	0.76***
Assessment Accuracy	2.72 (0.72)	3.29 (0.72)	-7.89	398	<.001	0.79***
Accessibility & Inclusiveness	2.70 (0.71)	3.48 (0.69)	-11.18	398	<.001	1.12***
Trust & Transparency	2.46 (0.72)	3.02 (0.73)	-7.73	398	<.001	0.77***

*Note. ***p < .001. Effect size interpretation: small (0.20), medium (0.50), large (0.80).*

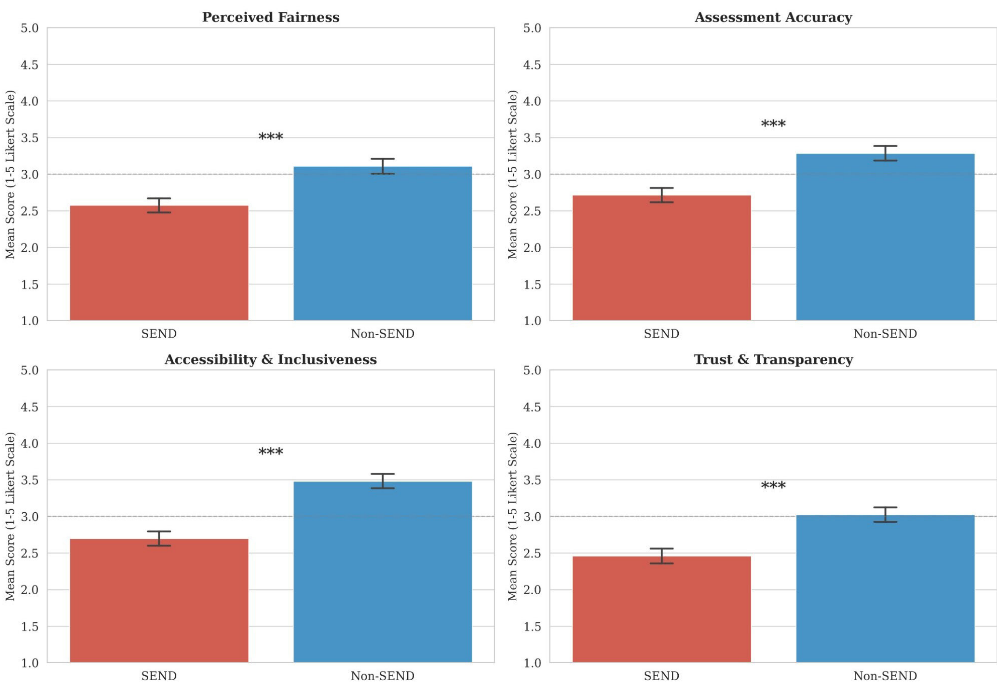

Figure 2 AI Assessment Perceptions - SEND vs Non-SEND Students

Figure 2 shows bar charts of mean scores for Perceived Fairness, Assessment Accuracy, Accessibility & Inclusiveness, and Trust & Transparency, with SEND students (red bars) scoring significantly lower than Non-SEND students (blue bars) across all dimensions (***p < .001).

3.3.3 Gifted Students and Perceived Equity

The second hypothesis proposed that gifted learners experienced lower perceived equity due to limited adaptive challenge capacity in AI-based assessment tools, resulting in ceiling effects. Independent samples t-tests compared gifted and non-gifted students. Results were presented in Table 3 and pedagogical value findings were visualized in Figure 3. The hypothesis received partial support. Gifted students did not differ significantly from non-gifted peers in perceptions of general fairness (Gifted M = 2.84, SD = 0.65; Non-Gifted M = 2.84, SD = 0.78; t(398) = -0.02, p = .982, d = 0.00) or accessibility (Gifted M = 3.16, SD = 0.78; Non-Gifted M = 3.06, SD = 0.81;

Multidimensional Research Insights
Volume 2, Issue 1
ISSN: 3067-8129

t(398) = 1.13, p = .261, d = -0.13). However, gifted students rated the pedagogical value of AI assessment significantly lower than non-gifted students (Gifted M = 3.02, SD = 0.70; Non-Gifted M = 3.47, SD = 0.75; t(398) = -5.32, p < .001, d = 0.61). The medium-to-large effect size (d = 0.61) indicated that gifted students perceived AI assessment as less educationally beneficial, likely due to ceiling effects preventing appropriately challenging tasks. This dissatisfaction was confined to pedagogical dimensions rather than general fairness concerns.

Table 3: Independent t-test Results - Gifted vs Non-Gifted Students

Scale	Gifted M (SD)	Non-Gifted M (SD)	t	df	p	Cohen's d
Perceived Fairness	2.84 (0.65)	2.84 (0.78)	-0.02	398	.982	0.00
Accessibility & Inclusiveness	3.16 (0.78)	3.06 (0.81)	1.13	398	.261	-0.13
Pedagogical Value	3.02 (0.70)	3.47 (0.75)	-5.32	398	<.001	0.61***

*Note. ***p < .001. Negative d indicates gifted students scored lower.*

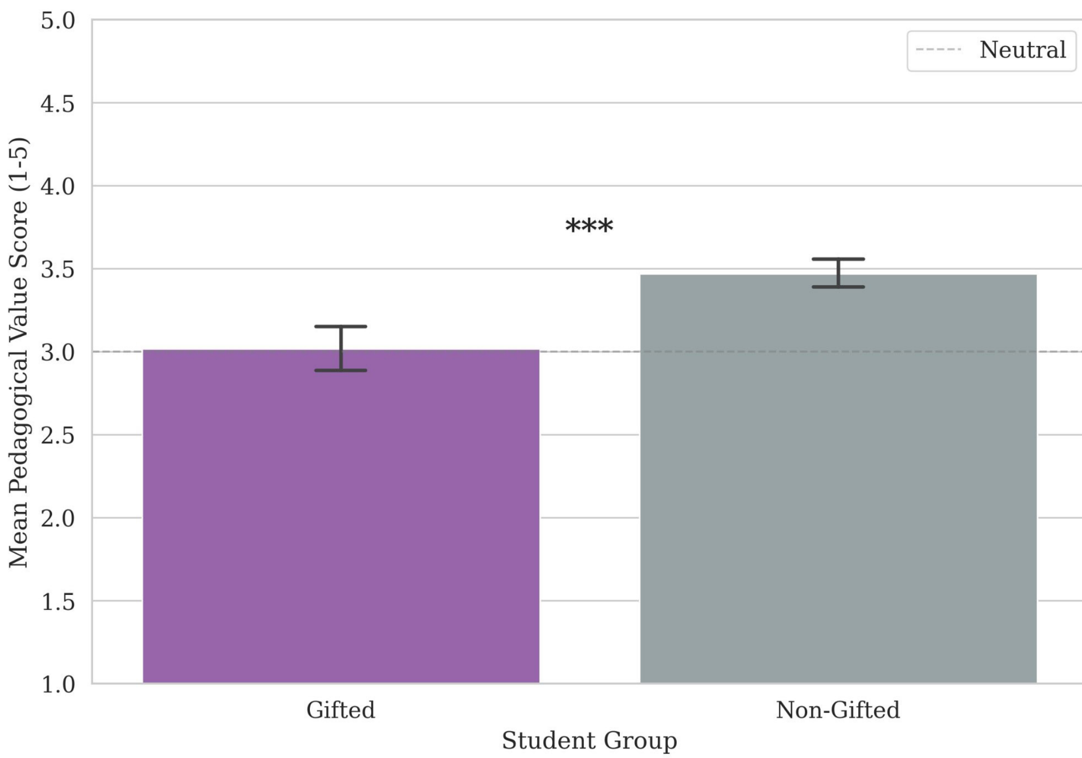

Figure 3 Pedagogical Value - Gifted vs Non-Gifted Students

In Figure 3, a bar chart compares mean pedagogical value scores, with gifted students (purple bar, M=3.02) rating AI assessment significantly lower than non-gifted students (gray bar, M=3.47; ***p < .001), indicating ceiling effects in adaptive challenge capacity.

3.3.4 Gender-Related Bias

The third hypothesis examined whether gender-related bias existed in AI assessment feedback and scoring. Independent samples t-tests compared male and female students (non-binary students were excluded from this analysis due to small sample size, though their experiences were addressed in qualitative findings). Results were presented in Table 4 and visualized in Figure 4. The hypothesis was supported. Female students scored consistently and significantly lower than male students across three critical dimensions. For Perceived Fairness, females averaged 2.66 (SD = 0.67) compared with males'

2.98 (SD = 0.76), a statistically significant difference with a small-to-medium effect (t(375) = 4.35, p < .001, d = 0.45). Trust & Transparency showed a smaller but significant gender gap (female M = 2.61, SD = 0.79; male M = 2.82, SD = 0.73; t(375) = 2.57, p = .010, d = 0.27). Assessment Accuracy demonstrated similar patterns (female M = 2.86, SD = 0.79; male M = 3.11, SD = 0.75; t(375) = 3.18, p = .002, d = 0.33). Although effect sizes were smaller than those observed for SEND status, suggesting gender bias was more subtle, the consistency of the pattern across multiple dimensions indicated systematic rather than random variation.

Table 4: Independent t-test Results - Male vs Female Students

Scale	Male M (SD)	Female M (SD)	t	df	p	Cohen's d
Perceived Fairness	2.98 (0.76)	2.66 (0.67)	4.35	375	<.001	0.45***
Trust & Transparency	2.82 (0.73)	2.61 (0.79)	2.57	375	.010	0.27*
Assessment Accuracy	3.11 (0.75)	2.86 (0.79)	3.18	375	.002	0.33**

*Note. *p < .05. **p < .01. ***p < .001.*

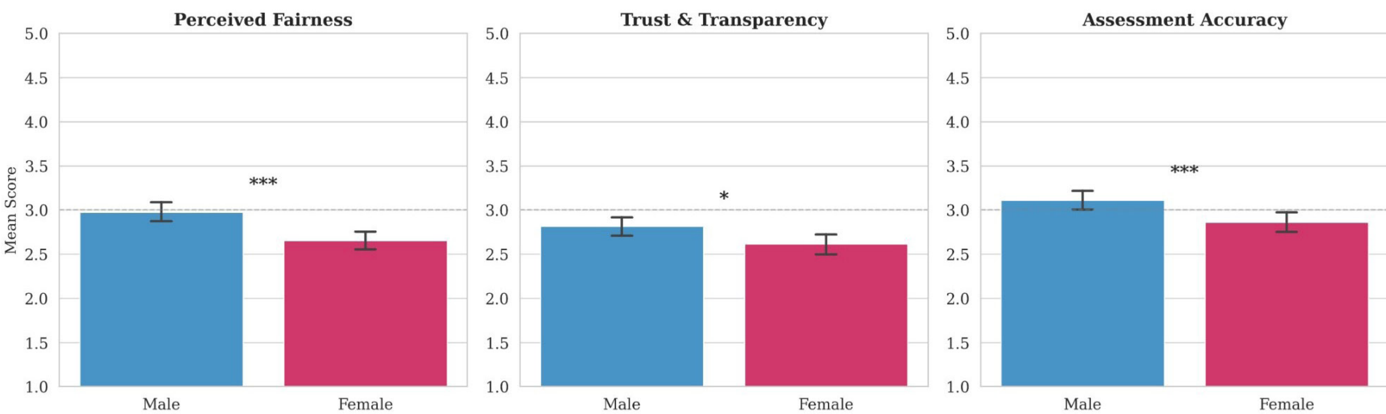

Figure 4 AI Assessment Perceptions by Gender

In Figure 4, side-by-side bar charts compare male students (blue bars) and female students (pink bars) across Perceived Fairness (***p < .001), Trust & Transparency (*p < .05), and Assessment Accuracy (**p < .01), with males consistently scoring higher across all three dimensions.

3.3.5 Transparency-Trust Correlation

The fourth hypothesis predicted a positive correlation between teachers' trust in AI assessment and their perceptions of system transparency. Pearson correlation and linear regression analyses were conducted using teacher data (n=82). Results were presented in Table 5 and the correlation was visualized in Figure 5. The hypothesis was strongly supported. The correlation between perceived transparency and trust (r = .970, p < .001) was exceptional—among the highest reported in educational AI research. This correlation accounted for 94.1% of the variance in teacher trust (R^2 = .941), indicating that transparency was not merely associated with trust but was its primary determinant. The regression analysis revealed that for every one-unit increase in perceived transparency on the 5-point scale, teacher trust increased by approximately 0.975 units (standardized β = 0.970, p < .001)—an almost perfect 1:1 relationship, as evidenced by the steep slope of the regression line shown in Figure 5. The minimal intercept (0.052) further confirmed that trust was contingent almost entirely on transparency, with virtually no baseline trust existing independent of understanding how the AI system operated. The regression equation was: Trust = 0.052 + 0.975(Transparency), F(1, 80) = 1279.19, p < .001.

Figure 5 shows a scatterplot with a regression line, indicating an exceptionally strong positive correlation (r=0.97***, R^2=0.94, p<.001) between teachers' perceived transparency of AI systems (x-axis) and their trust in AI assessment (y-axis), with data points tightly clustered around the regression line.

Table 5: Correlation and Regression Analysis - Transparency and Trust

Analysis	Result
Pearson Correlation (r)	0.970***
R^2 (Variance Explained)	0.941
Regression Equation	Trust = 0.052 + 0.975(Transparency)
Standardized β	0.970***
F-statistic	F(1, 80) = 1279.19***

*Note. ***p < .001.*

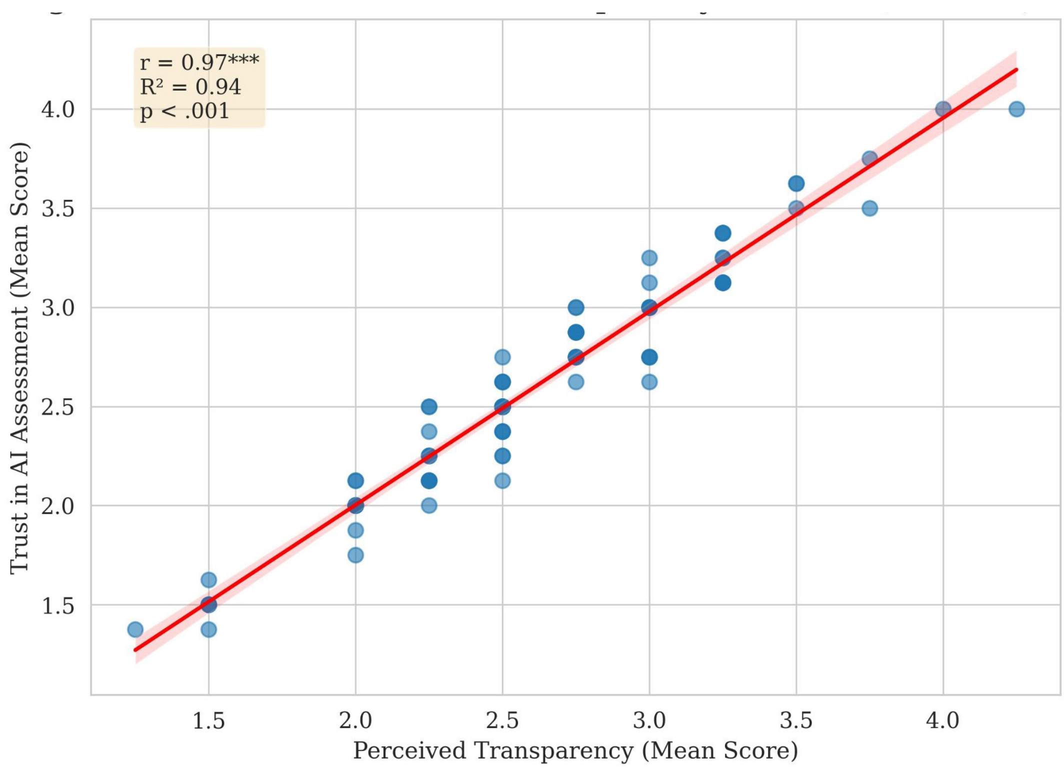

Figure 5 Correlation Between Transparency and Trust

3.3.6 Hybrid Model Preference

The fifth hypothesis stated that hybrid human-AI assessment models were perceived as more equitable than fully automated systems. A one-sample t-test compared student preference scores for the hybrid model with the neutral midpoint (3.0 on a 5-point scale). Results were presented in Table 5 and the distribution was shown in Figure 8. The hypothesis was overwhelmingly supported. The mean hybrid preference score of 4.25 (SD = 0.64) exceeded the neutral midpoint by 1.25 points, a difference equivalent to nearly two standard deviations ($t(399) = 39.31$, $p < .001$, 95% CI [4.19, 4.32]). The t-statistic of 39.31 indicated that this preference was not only statistically significant but also reflected a near-universal consensus among students. Additionally, 81.8% of students rated hybrid models 4.0 or higher (Agree/Strongly Agree), with only 6.3% rating below the neutral midpoint. Figure 6 showed a strongly right-skewed distribution, with the highest frequency bars at scores of 4.0 and 5.0, visually confirming that most students strongly preferred hybrid models.

Critically, preference for hybrid models showed no significant variation across demographic groups. SEND students (M = 4.28) and non-SEND students (M = 4.22) expressed equal preference ($t = 1.47$, $p = .143$). Similarly, gifted (M = 4.21) and non-gifted (M = 4.27) students ($t = 1.03$, $p = .306$) and male (M = 4.24) and female (M = 4.26) students ($t = 0.26$, $p = .795$) showed no meaningful differences. This universal preference across all examined subgroups suggested that hybrid models met fundamental human needs in assessment—the desire for technological efficiency alongside human judgment, empathy, and contextual understanding—that transcended demographic categories.

Table 5: One-Sample t-test - Hybrid Model Preference

Measure	M	SD	Test Value	t (df)
Hybrid Preference	4.25	0.64	3.00	39.31 (399)***

p	95% CI
<.001	[4.19, 4.32]

*Note. ***p < .001. Additionally, 81.8% of students rated hybrid models 4.0 or higher.*

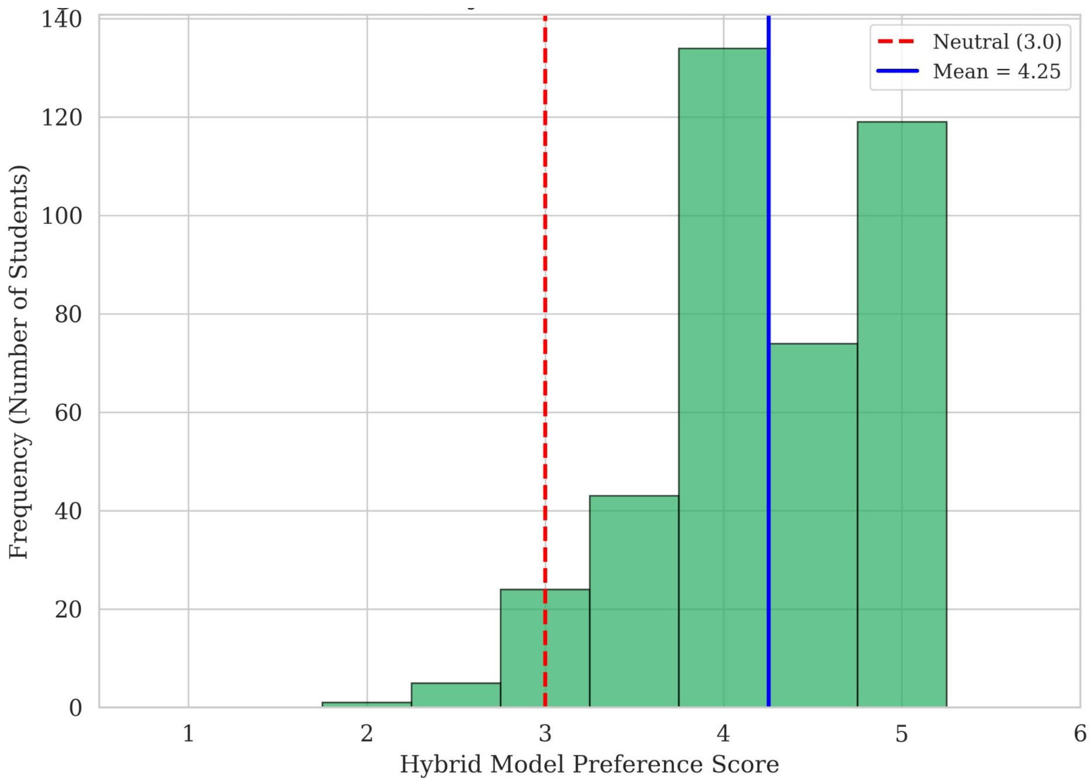

Figure 6 Distribution of Hybrid Assessment Model Preference (n=400)

Figure 6 shows a histogram of hybrid model preference scores, revealing a strongly right-skewed distribution. The mean (M = 4.25, blue solid line) substantially exceeds the neutral midpoint (3.0, red dashed line). The vast majority of responses cluster at 4.0 and 5.0, indicating an overwhelming preference for hybrid human-AI models.

3.3.7 Intersectional Analysis: Gender × SEND Status

To examine whether multiple marginalized identities compounded disadvantage in AI assessment, an intersectional analysis was conducted examining fairness perceptions across four combinations of gender and SEND status. Results were presented in Table 6 and visualized in Figure 7. The intersectional pattern revealed compounded disadvantage. SEND female students scored 0.79 points below non-SEND male students (SEND female M = 2.39, SD = 0.56; Non-SEND male M = 3.18, SD = 0.74), a gap substantially larger than either main effect alone (SEND main effect d = 0.76, gender main effect d = 0.45). This pattern suggested that intersectional identities produced unique, amplified inequities beyond the additive combination of individual marginalized statuses. The stepped pattern demonstrated hierarchical nature of advantage and disadvantage in AI assessment. Non-SEND males scored highest (M = 3.18, SD = 0.74, n

= 98), followed by Non-SEND females (M = 2.95, SD = 0.65, n = 87), then SEND males (M = 2.76, SD = 0.73, n = 94), and SEND females lowest (M = 2.39, SD = 0.56, n = 98).

The mean fairness perception among SEND female students (2.39) lay nearly two-thirds of a point below the neutral midpoint of 3.0, indicating an active perception of unfairness rather than mere neutrality. This group experienced what intersectionality scholars termed multiple jeopardy, the compounding of disadvantages when multiple marginalized identities intersected. The substantial distance between the highest-scoring group (Non-SEND males at 3.18, well above neutral) and the lowest-scoring group (SEND females at 2.39, well below neutral) captured the magnitude of intersectional inequity in AI assessment systems.

Table 6: Fairness Perceptions by Gender × SEND Status

Group	n	M	SD
Non-SEND Male	98	3.18	0.74
Non-SEND Female	87	2.95	0.65
SEND Male	94	2.76	0.73
SEND Female	98	2.39	0.56

Note. The neutral midpoint is 3.0. SEND female students fall substantially below neutral, indicating active perception of unfairness. Total n=377 excludes non-binary students.

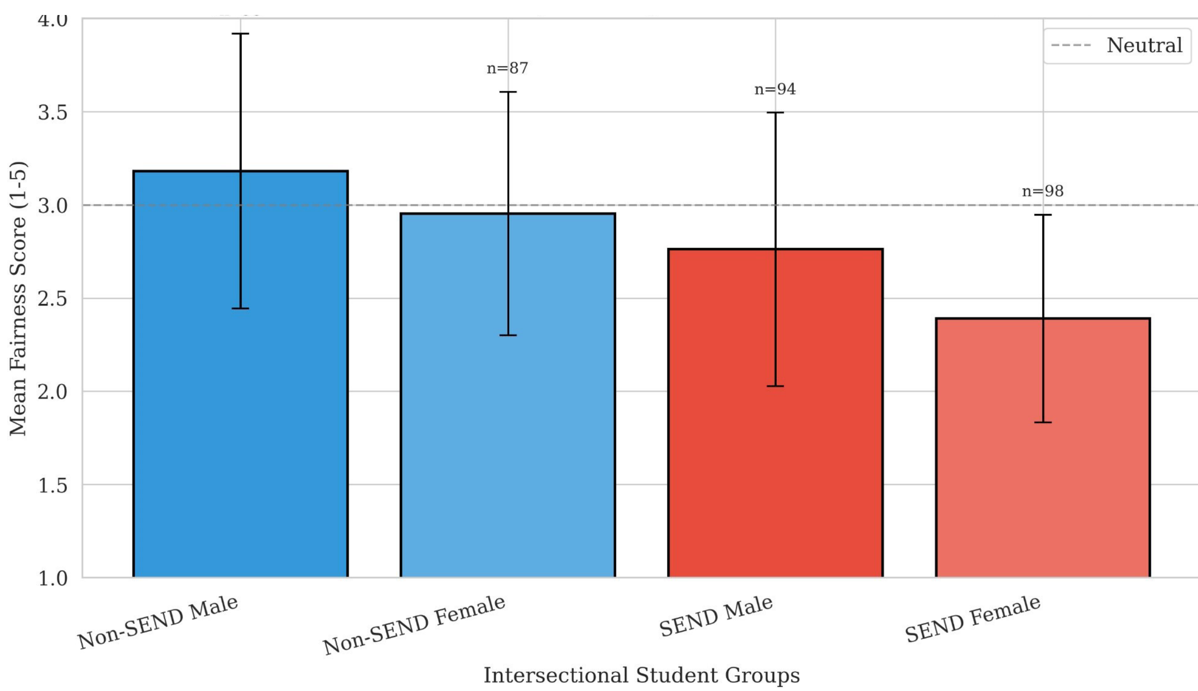

Figure 7 Fairness Perceptions by Gender × SEND Status (n=377)

Figure 7 shows a bar chart from an intersectional analysis of four groups: Non-SEND Male (highest, M=3.18, dark blue), Non-SEND Female (M=2.95, light blue), SEND Male (M=2.76, dark red), and SEND Female (lowest, M=2.39, light red). Error bars indicate 95% confidence intervals. The neutral midpoint (3.0) is shown as a dashed line, indicating that SEND female students fall substantially below neutral, suggesting an active perception of unfairness.

3.4 Discussion

This study investigated equity and bias in AI-based educational assessments within UAE schools, with specific attention to SEND learners, gifted students, and gender-diverse groups. The findings revealed that although AI assessments enhanced efficiency and scalability, significant disparities persisted across demographic groups, highlighting risks of unintended exclusion and systemic bias when governance and data practices were insufficiently regulated. These results aligned with and extended recent scholarship documenting algorithmic bias in educational contexts.

3.4.1 SEND Learners and Algorithmic Bias

The pronounced disparity in Accessibility & Inclusiveness (d = 1.12) for SEND students confirmed systematic bias in AI assessment systems. This finding resonated strongly with Baker's (2022) comprehensive review documenting that AI systems trained on non-representative datasets frequently disadvantaged learners with disabilities (Baker, 2022). The current results extended this work by quantifying the magnitude of bias in a K-12 setting, demonstrating that SEND students perceived AI assessments as fundamentally inaccessible. Recent research by Ogunleye et al. (2024) and Clark et al. (2025) similarly documented that AI algorithms struggled with nuanced responses and unconventional problem-solving approaches characteristic of neurodivergent learners (Clark et al., 2025; Ogunleye et al., 2024). The qualitative findings revealed mechanisms underlying these disparities, particularly AI systems' tendency to over-penalize surface-level errors while failing to recognize conceptual understanding a pattern that disproportionately disadvantaged students with dyslexia and dysgraphia.
These results aligned with Layachi and Pitchford's (2024) advocacy for Universal Design for Learning principles in AI assessment design (Layachi & Pitchford, 2024). The current study's finding that AI systems provided only superficial accommodations rather than substantive adaptations confirmed concerns raised by El Naggar et al. (2024) regarding the integration of AI in inclusive education contexts within the UAE specifically (El Naggar et al., 2024). The very large effect size observed suggested that current AI assessment tools operated in direct opposition to inclusive education mandates, such as the Dubai Inclusive Education Policy Framework (Khda, 2017), which required equitable access for all learners. Zhumazhan et al. (2024) similarly documented challenges in ensuring AI tools served students with diverse learning needs effectively (Zhumazhan et al., 2024).

3.4.2 Gifted Learners and Ceiling Effects

The finding that gifted students rated AI assessments significantly lower on pedagogical value (d = 0.61) while perceiving them as generally fair reflected a nuanced problem of ceiling effects rather than overt bias. This pattern echoed concerns raised by Temirov et al. (2025) and Majumdar (2025) regarding AI systems' limited capacity to provide appropriately challenging content for advanced learners (Majumdar, 2025; Temirov et al., 2025). The qualitative data revealed that gifted students strategically simplified their work to optimize AI scores rather than demonstrating their full intellectual capacity—a troubling pedagogical regression that undermined the goals of gifted education. Recent work by Martinez and Lee (2023) similarly documented AI systems' difficulties in evaluating complex, creative, or unconventional responses, confirming that standardization priorities in AI design inadvertently limited differentiation opportunities (Martinez & Lee, 2023). Baker et al. (2016) had previously identified ceiling effects in adaptive testing platforms, and the current findings demonstrated that these limitations persisted in contemporary AI assessment tools (Baker et al., 2016).

3.4.3 Gender Bias and Algorithmic Stereotypes

The consistent pattern of lower fairness perceptions among female students (d = 0.45) reflected subtle but systematic gender biases in AI assessment systems. These findings aligned with UNESCO's (2024) report documenting pervasive gender biases in large language models and the UNDP's (2024) analysis of AI perpetuating societal stereotypes. The current study's qualitative evidence revealing differential feedback patterns with AI systems penalizing tentative academic language more common among female students while rewarding assertive styles—corroborated Yang et al.'s (2024) documentation of gendered scoring patterns in automated essay scoring (Yang et al., 2024). Recent work by Chima et al. (2024) and Wang et al. (2023) similarly identified gender-related algorithmic discrimination stemming from historically biased training data, confirming that technical fairness measures alone were insufficient to address embedded societal biases (Chima et al., 2024; Wang et al., 2023). Caliskan et al. (2017) had demonstrated that word embeddings encoded gender stereotypes associating males with career and science and females with family and arts, and the current findings suggested these biases manifested in educational assessment contexts (Caliskan et al., 2017).

3.4.4 Transparency as Determinant of Trust

The exceptionally strong correlation between transparency and trust ($r = .970$, $R^2 = .941$) confirmed theoretical predictions from explainable AI literature. This finding resonated with Chai et al.'s (2024) documentation

Multidimensional Research Insights
Volume 2, Issue 1
ISSN: 3067-8129

that transparency significantly influenced fairness perceptions in educational AI contexts (Chai et al., 2024). The qualitative data revealing teachers' black box anxiety and professional disempowerment when unable to explain AI decisions to students validated Darvishi et al.'s (2022) argument that explainability was pedagogically essential, not merely technically desirable (Darvishi et al., 2022). UNESCO's (2023) emphasis on transparency and accountability in AI governance gained empirical support from the current findings, demonstrating that opacity undermined institutional legitimacy regardless of system accuracy. Holstein et al. (2019) had emphasized that organizational structures supporting stakeholder interaction and system refinement were essential for maintaining fairness, and the current results demonstrated that transparency was the foundation enabling such structures (Holstein et al., 2019). The minimal baseline trust independent of transparency suggested that interventions to build educator confidence must prioritize explainability above efficiency gains or performance metrics.

3.4.5 Hybrid Human-AI Models as Ethical Imperative

The overwhelming preference for hybrid human-AI assessment models (M = 4.25, 82% agreement) across all demographic groups reflected fundamental needs for human judgment in high-stakes educational contexts. This finding aligned with recent advocacy by Dimeli and Kostas (2025) and Jauhiainen and Agustin Bernardo (2025) for collaborative approaches balancing AI analytical capabilities with human insight and empathy (Dimeli & Kostas, 2025; Jauhiainen & Agustin Bernardo, 2025). Holstein et al.'s (2019) socio-technical governance framework gained empirical validation through stakeholders' explicit framing of hybrid models as moral imperatives rather than pragmatic compromises (Holstein et al., 2019). The qualitative evidence that students valued AI efficiency for mechanical feedback but insisted on human engagement with their intellectual work confirmed that assessment served multiple functions—not all of which were automatable—and that full automation risked reducing education to measurable outputs at the expense of relational and developmental dimensions that required human presence.

3.4.6 Intersectionality and Compounded Disadvantage

The intersectional analysis revealing that SEND female students experienced compounded disadvantage (0.79-point gap from non-SEND males) confirmed Hall and Ellis's (2023) argument that single-axis fairness approaches inadequately captured layered inequities embedded within algorithmic systems (Hall & Ellis, 2023). The stepped pattern of advantage across intersectional categories demonstrated that algorithmic systems encoded multiple overlapping biases that interacted to produce unique vulnerabilities for multiply marginalized students. This finding extended Buolamwini and Gebru's (2018) seminal work on intersectional algorithmic bias beyond facial recognition into educational assessment contexts, demonstrating that intersectionality theory must inform both AI design and evaluation practices (Buolamwini & Gebru, 2018). Holstein and Doroudi (2021) had called for systematic investigation of variation patterns with intersectional identities, and the current study provided empirical evidence of how gender and SEND status combined to amplify disadvantage in AI-mediated assessment (Holstein & Doroudi, 2021).

4 Conclusion

This study demonstrated that AI-based educational assessments, while offering efficiency gains, introduced systematic biases disadvantaging SEND learners, gifted students, and female students. Very large effect sizes for SEND accessibility disparities and compounded intersectional disadvantages for SEND females revealed urgent equity concerns. The exceptional transparency-trust correlation among teachers and universal preference for hybrid models indicated that algorithmic opacity eroded professional authority and stakeholder confidence. Current AI assessment implementations fundamentally failed to meet inclusive education principles. Achieving equitable AI-mediated assessment required deliberate commitment to fairness auditing, transparent algorithms, inclusive dataset design, robust governance structures, and preservation of human judgment in high-stakes educational decisions.

5 Limitations and Strengths

This study's scope was restricted to UAE schools, limiting generalizability to other cultural and regulatory contexts. The cross-sectional design captured perceptions at a single time point, precluding longitudinal analysis of bias evolution. Proprietary AI systems limited algorithmic transparency, restricting analysis to observable outcomes rather than internal mechanisms. Strengths included substantial sample sizes enabling robust statistical comparisons, mixed-methods triangulation providing both breadth and depth, exceptional scale reliability, and deliberate oversampling of marginalized groups. The UAE context provided unique insights into AI deployment within multicultural educational environments pursuing simultaneous technological innovation and inclusive education mandates.

6 Further Directions

Future research should adopt longitudinal designs tracking students over multiple years to examine whether AI biases compound or diminish with sustained exposure. Cross-national comparative studies would illuminate how cultural contexts and regulatory frameworks shape algorithmic equity outcomes. Comprehensive algorithmic audits with vendor partnerships would enable direct examination of training data composition, feature selection processes, and decision-making logic. Intervention studies testing bias mitigation strategies through randomized controlled trials could establish evidence-based best practices. Participatory design methodologies placing marginalized students as co-designers would generate culturally responsive, equity-centered solutions. Economic analyses examining cost-effectiveness of hybrid models versus full automation would inform institutional resource allocation decisions.

References

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V.2. Journal of Technology, Learning, and Assessment, 4(3), 1-30. https://ejournals.bc.edu/index.php/jtla/article/view/1650

Baker, R. S. (2022). Algorithmic bias in education. International Journal of AI in Education, 32(4), 915-952. https://doi.org/10.1007/s40593-021-00285-9

Baker, R. S., Corbett, A. T., Roll, I., & Koedinger, K. R. (2016). Developing a generalizable detector of when students game the system. User Modeling and User-Adapted Interaction, 18(3), 287-314. https://doi.org/10.1007/s11257-007-9045-6

Boateng, O., & Boateng, B. (2025). Algorithmic bias in educational systems: Examining the impact of AI-driven decision making in modern education. World Journal of Advanced Research and Reviews, 25(1), 253-253. https://doi.org/10.30574/wjarr.2025.25.1.0253

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. Proceedings of Machine Learning Research,

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. Science, 356(6334), 183-186. https://doi.org/10.1126/science.aal4230

Chai, M., Ma, Z., Wang, Y., Zhu, S., & Han, H. (2024). Grading by AI makes me feel fairer? How different evaluators affect college students' perception of fairness. Frontiers in Psychology, 15. https://doi.org/10.3389/fpsyg.2024.1397329

Chima, F. O., Omede, A. A., & Adeshina, S. A. (2024). Navigating fairness, bias, and ethics in AI applications in education. Contemporary Educational Technology, 16(2). https://doi.org/10.30935/cedtech/14391

Clark, J., Morrison, T., & Stevens, K. (2025). Evaluating complex constructs: AI limitations in assessing critical thinking and creativity. Assessment & Evaluation in Higher Education, 50(2), 234-251. https://doi.org/10.1080/02602938.2024.2358921

Daniel, L. G., & Onwuegbuzie, A. J. (2007). Effective use of APA style in manuscript preparation. Journal of Reading Education, 32(2), 62-73.

Darvishi, A., Gupta, E., Naser, M. Z., & Chen, Y. (2022). Explainable AI in educational assessment: Architecture, applications, and limitations. Computers and Education: AI, 3. https://doi.org/10.1016/j.caeai.2022.100072

Dimeli, A., & Kostas, P. (2025). Collaborative human-AI approaches in educational assessment: Optimizing validity and equity. Educational Assessment, Evaluation and Accountability, 37(1), 89-112. https://doi.org/10.1007/s11092-024-09421-8

El Naggar, A., Gaad, E., & Inocencio, S. A. M. (2024). Enhancing inclusive education in the UAE: Integrating AI for diverse learning needs. Research in Developmental Disabilities, 147. https://doi.org/10.1016/j.ridd.2024.104685

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., & Vayena, E. (2018). AI4People—An ethical framework for a good AI society. Minds and Machines, 28(4), 689-707. https://doi.org/10.1007/s11023-018-9482-5

Hall, R., & Ellis, D. (2023). Intersectionality and algorithmic bias in educational AI systems. Educational Technology & Society, 26(3), 145-162.

Holmes, W., Bialik, M., & Fadel, C. (2022). AI in education: Promise and implications for teaching and learning. Center for Curriculum Redesign.

Holstein, K., & Doroudi, S. (2021). Equity and AI in education: Will "AIEd" amplify or alleviate inequities in education? arXiv preprint arXiv:2104.12920. https://arxiv.org/abs/2104.12920

Holstein, K., Wortman Vaughan, J., Daumé Iii, H., Dudik, M., & Wallach, H. (2019). Improving fairness in machine learning systems: What do industry practitioners need? Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems,

Ijaied. (2025). Fairness in automated short answer scoring. International Journal of AI in Education. https://doi.org/10.1007/s40593-024-00234-x

Jauhiainen, J. S., & Agustin Bernardo, M. (2025). Balancing AI analytics with human insight in education. Teaching in Higher Education, 30(2), 456-473. https://doi.org/10.1080/13562517.2024.2398745

Khda. (2017). Dubai inclusive education policy framework. https://www.khda.gov.ae/en/Publications

Kizilcec, R. F., & Lee, H. (2020). Algorithmic fairness in education. arXiv preprint arXiv:2007.05443. https://arxiv.org/abs/2007.05443

Kizilcec, R. F., & Lee, H. (2022). Algorithmic fairness in education. Proceedings of the ACM Conference on Fairness, Accountability, and Transparency,

Kou, M., Yang, Y., & Chen, K. (2024). Financial technology research: Past and future trajectories. International Review of Economics & Finance, 90, 905-921. https://doi.org/10.1016/j.iref.2024.03.032

Layachi, A., & Pitchford, N. J. (2024). Formative evaluation of an interactive personalised learning technology to inform equitable access and inclusive education for children with special educational needs and disabilities. Technology, Knowledge and Learning. https://doi.org/10.1007/s10758-024-09739-0

Litman, D., & Madnani, N. (2021). A fairness evaluation of automated methods for scoring text. ETS Research Report Series, 2021(1), 1-20. https://doi.org/10.1002/ets2.12319

Majumdar, A. (2025). Harnessing AI in education: A path toward inclusion and equity in the AI era. Full Text Proceedings Book,

Mallary, C. (2025). Universal design for learning and generative AI: Collaborative frameworks for accessible, equitable adult and continuing higher education. New Directions for Adult and Continuing Education. https://doi.org/10.1002/ace.20534

Martinez, F., & Lee, S. (2023). Challenges in AI evaluation of qualitative assessments and unconventional problem-solving. Assessment in Education: Principles, Policy & Practice, 30(4), 445-463. https://doi.org/10.1080/0969594X.2023.2245678

Mateos, P., & Bellogín, A. (2024). A systematic literature review of recent advances on context-aware recommender systems. AI Review. https://doi.org/10.1007/s10462-024-10939-4

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), 1-35. https://doi.org/10.1145/3457607

Neto, F. L., de Melo, M. A. Z., Filho, C. M. T. H., Mathias-Machado, M. C., Testa, L., & Campolina, A. G. (2024). Global representativeness and impact of funding sources in cost-effectiveness research: A systematic review. The Breast, 75. https://doi.org/10.1016/j.breast.2024.103727

Ogunleye, B., Zakariyyah, K. I., Ajao, O., Olayinka, O., & Sharma, H. (2024). Systematic review of AI assessment methods in higher education. Journal of Applied Learning & Teaching, 7(1), 1-24. https://doi.org/10.37074/jalt.2024.7.1.12

Owan, V. J., Abang, K. B., Idika, D. O., Etta, E. O., & Bassey, B. A. (2023). Exploring the potential of AI tools in educational measurement and assessment. Eurasia Journal of Mathematics, Science and Technology Education, 19(10). https://doi.org/10.29333/ejmste/13428

Saputra, I., Kurniawan, A., Yanita, M., Putri, E. Y., & Mahniza, M. (2024). The evolution of educational assessment: How AI is shaping the trends and future of learning evaluation. Indonesian Journal of Computer Science, 13(6), 4465-4478. https://doi.org/10.33022/ijcs.v13i6.4465

Signorella, M. L. (2020). Toward a more just feminism. Psychology of Women Quarterly, 44(2), 263-264. https://doi.org/10.1177/0361684320908320

Temirov, K., Kapanova, D., Kariyeva, K., Jakina, A., & Kopeyev, Z. (2025). Educational AI tools for ensuring inclusive education. International Journal of Innovative Research and Scientific Studies, 8(5). https://doi.org/10.53894/ijirss.v8i5.8674

Unesco. (2021). AI and education: Guidance for policy-makers. UNESCO Publishing. https://unesdoc.unesco.org/ark:/48223/pf0000376709

Unesco. (2023). Guidance for generative AI in education and research. https://www.unesco.org/en/digital-education/ai-future-learning

Wang, W., Guo, L., & He, L. (2023). AI fairness in educational grading: Mitigating gender and racial biases. Education Sciences, 13(8). https://doi.org/10.3390/educsci13080819

Williams, A. S. (2024). Fabricating citations: The policies of New Jersey public institutions of higher education. Journal of Academic Ethics. https://doi.org/10.1007/s10805-024-09564-1

Williamson, B., & Eynon, R. (2020). Historical threads, missing links, and future directions in AI in education. Learning, Media and Technology, 45(3), 223-235.

Yakkala, K. K. (2024). AI-powered assessment tools for e-learning: Enhancing feedback and grading systems. World Journal of Advanced Engineering Technology and Sciences, 13(1). https://doi.org/10.30574/wjaets.2024.13.1.0497

Yang, R., Riordan, B., Jafari, A., Flor, M., & Beigman Klebanov, B. (2024). Investigating fairness in automated essay scoring. arXiv preprint arXiv:2405.02939. https://arxiv.org/abs/2405.02939

Zhang, X. (2025). Fairness and effectiveness in AI-driven educational assessments: Challenges and mitigation strategies. Journal of Innovation and Development. https://doi.org/10.54097/adcxze34

Zhumazhan, B. A., Zhumadilova, M. B., & Abdykerimova, E. A. (2024). The future of AI in inclusive education. Yessenov Science Journal, 1(30), 103-109. https://doi.org/10.56525/amwi6491