Policy-Sensitive Fairness Evaluation in Automated Scoring of Clinical Communication

Saed Rezayi[1], Le An Ha[2], Victoria Yaneva[1]
Polina Harik[1], Janet Mee[1], Jason Snyder[1]

[1]NBME [2]Ho Chi Minh City University of Foreign Languages

(srezayidemne,vyaneva,pharik,jmee,jsnyder)@nbme.org anhl@huflit.edu.vn

Abstract

This study examines automated scoring fairness in a formative assessment context: the automated evaluation of medical students' communication skills. Building on the premise that definitions of fairness are value-dependent, we investigate how conclusions about group differences may vary under different weighting schemes for false positives (FPs) and false negatives (FNs). Results show that when errors are treated symmetrically, no statistically significant differences are observed across demographic groups based on race or gender. This pattern remains stable when error weights are varied, with no consistent or robust disparities emerging. A small number of isolated differences appear under moderate FN weighting. Overall, the findings suggest that fairness conclusions in this setting are relatively robust to variations in error weighting. At the same time, the study highlights the importance of making value assumptions explicit when evaluating automated scoring systems, particularly in formative contexts where error trade-offs carry pedagogical implications for feedback, learner engagement, and educational equity.

1 Introduction

Automated scoring systems are increasingly used in educational settings to evaluate constructed responses at scale (Flores et al., 2025). They can reduce scoring burden and support timely feedback, but they also raise an important question: does the system behave differently across demographic groups in ways that reflect construct-irrelevant factors? This concern is especially important in cases where scores may affect learning opportunities, consequential decisions, and perceptions of competence.

Fairness is an inherent concern in algorithmic classification, and a consistent finding across the literature is that fairness is not a single, unified concept. Notable papers such as *"On the (im)possibility of fairness"* (Friedler et al., 2021) showcase how different notions of fairness reflect different underlying value systems, and no single classification system can satisfy all of them simultaneously.

This challenge is equally present in research on automated scoring in education (Johnson and McCaffrey, 2023). In this field, a fair test is one in which differences in scores among test-takers reflect only variations in the skills defined by the construct. Systematic score differences between non-random groups that arise from factors unrelated to the construct (i.e., construct-irrelevant factors) may signal potential unfairness in the assessment (Litman et al., 2021; Schaller et al., 2024; Andersen et al., 2025). To illustrate these competing perspectives, Loukina et al. (2019) distinguish three aspects of fairness: (1) whether scores are equally accurate across groups, (2) whether scores systematically differ from human scores for certain groups, and (3) whether such discrepancies persist when individuals have comparable ability. Empirical findings corroborate that fully satisfying all criteria simultaneously is rarely feasible.

While prior work on fairness in automated scoring has primarily focused on *summative assessment*, particularly for essay scoring and short-answer grading (Johnson et al., 2022), this paper examines fairness in *formative assessment* and in a different domain: the automated evaluation of medical students' communication skills with patients.

When discussing fairness in automated scoring in learning tools, there could be different value systems that govern the definition and subsequent evaluation of fairness. In the present setting, our automated scoring system makes binary decisions about whether specific behaviors are present in a learner response. In this setting, one key consideration is the relative importance of false positives (FPs) versus false negatives (FNs), as discussed in several key studies (Dey et al., 2024; Ernst et al.,

Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications, pages 574–580

2025). In formative assessment, these trade-offs have pedagogical consequences: FNs may undermine learner confidence, while FPs can obscure learning gaps. Some value systems treat errors symmetrically, prioritizing overall accuracy; others emphasize minimizing FPs, ensuring positive evaluations reflect demonstrated competence; still others prioritize minimizing FNs to avoid discouraging learners or overlooking emerging skills. For example, Li et al. (2023) examine the effects of FPs (incorrect answers marked correct) and FNs (correct answers marked incorrect) in an AI autograder for short-answer tasks. The findings show that FPs can harm learning, largely because students fail to detect them and disengage after receiving positive feedback. FNs have more variable effects, with evidence that deeper engagement can mitigate their impact. In another study, Dey et al. (2024) find that the odds of addressing a FP feedback was 99% lower than addressing a FN feedback, representing significant missed opportunities for revision and learning. Thus, evaluation criteria reflect not only technical choices but also assumptions about which errors are most consequential for learning.

This study examines how fairness across demographic groups may vary under these different error-weighting approaches. If one type of error disproportionately affects certain learners, feedback may be misleading, impacting motivation and learning outcomes. Understanding these dynamics is crucial for designing automated scoring systems that are both accurate and educationally equitable. The paper makes the following original contributions:

- Examines fairness in automated scoring of medical students' communication skills in a formative context.
- Investigates the impact of error-weighting on fairness across demographic groups.
- Highlights the pedagogical implications of error asymmetries for feedback quality, learner motivation, and educational equity.

2 Data

Context: The context for this study is the Communication Learning Assessment (CLA) learning tool[1]. CLA is a formative, digital tool designed to support medical students in developing communication skills for effective interaction with patients. CLA presents learners with video-based patient scenarios that describe a clinical case. Each case ends with a question or statement from the patient. The learner can then record a verbal response of up to two minutes addressing the patient, where they strive to exemplify appropriate communication skills. The three main categories of skills practiced via these cases include 1) responding to emotions, 2) providing information, and 3) fostering the relationship. Within each of these categories, the learner is expected to demonstrate appropriate communication behaviors within the specific clinical case they are given, which are called *patient-centered behaviors (PCBs)*. For example, when responding to emotions in a case about a patient sharing discouragement that their weight loss journey has plateaued, an appropriate PCB might be *"I am sorry to hear this, it must be truly frustrating after all these efforts"*. Notably, these PCBs can vary widely among learners and cases, which makes PCB annotation and, subsequently, detection, a non-trivial task. In CLA, automated detection of case-specific PCBs is used to generate scores and those scores can form a basis of individualized feedback for learners.

Group	**Sample size (n)**	**PCB rate (%)**
Gender		
Female	837	61.4
Male	522	60.7
Non-binary	16	56.3
Other gender	8	87.5
race		
White/Caucasian	799	60.8
Asian	322	59.9
African American	150	63.3
Multiracial	56	58.9

Table 1: Dataset summary: subgroup sizes and annotation PCB prevalence ($N = 1{,}383$).

Learners: The learners in our dataset were 3rd and 4th year US medical students who had successfully passed the USMLE® Step 1 exam[2] and used the CLA tool for practicing their communication skills. The dataset includes self-reported gender and race for each learner.

Annotation: Responses from 223 learners across 8 clinical cases were annotated for PCBs, although not every learner responded to every case. Annotation was performed by a team of four NBME staff members trained in clinical communication assessment. Annotations were guided by a detailed rubric (see Appendix A), capturing

[1] https://www.nbme.org/cla

[2] https://www.usmle.org/step-exams/step-1

PCBs essential for effective physician-patient interactions, such as *acknowledgment of patient concerns, provision of clear explanations, demonstration of empathy,* and *reinforcement of positive behaviors*. This rubric included 25 unique PCBs, each associated exclusively with one of the 8 clinical cases, with each case containing between 1 and 4 distinct PCBs. Annotators were instructed to precisely identify text spans corresponding to each PCB by providing exact character-level indices. Negative samples for each PCB were systematically derived by listing all learner responses from the same clinical case that were not annotated as reflecting that specific PCB.

Annotators scored de-identified transcripts, were blind to learner demographics, and received explicit bias training. We therefore use human annotations as the reference standard for PCB presence, while recognizing that this does not rule out all sources of annotation bias.

Automated scoring system: The automated scoring of PCBs was performed by training ACTA (Rezayi et al., 2025) – an encoder-based transformer scoring model built on DeBERTa used operationally within CLA – on a cross-validated set of learner responses. Average binary F1 score for each PCB was 0.92 (min = 0.86, max = 0.97). Fairness was evaluated using out-of-fold predictions from 5-fold cross-validation on the annotated dataset.

Dataset: The resulting dataset consists of 1,383 records. Each record includes self-reported learner demographics, human PCB annotations, and ACTA predictions by learner response (encounter) and PCB type. Each prediction is assigned TP/TN/FP/FN labels after a careful error analysis by the annotation team. Each record includes the clinical communication text, a PCB label, and identifiers (`case_id`, `encounter_id`).

Table 1 reports subgroup sizes and human annotation prevalence (PCB rate). The largest groups are Female and White/Caucasian learners. Importantly, annotation PCB prevalence is nearly constant across high-volume groups (approximately 59–63%). Smaller groups (e.g., non-binary and other gender identities) have limited sample size and are therefore reported descriptively.

Looking at individual clinical cases (Table 4 in Appendix B), some cases show differences in PCB prevalence rates, but these differences are not statistically significant. In case 180, the PCB prevalence rate for the White/Caucasian group is around 10% lower than for the other two high-volume groups, while case 192 shows the reverse pattern. This suggests that differences in individual cases may reflect factors other than annotation bias.

3 Methodology

We use prediction scores from the operationally deployed ACTA model. ACTA produces binary PCB-level decisions for each clinical encounter (learner-case interaction), and a single encounter may contain multiple PCBs. Because these decisions are clustered within encounter, we use the encounter as the unit of analysis rather than treating PCB-level observations as independent.

For encounter e we define a policy-sensitive error score by summing weighted false negatives and false positives across all PCBs associated with the encounter's case:

$$L(e) \;=\; \sum_{c \in \mathcal{C}_{v(e)}} \left[w^{\mathrm{FN}}\, \mathrm{fn}_{e,c} \;+\; w^{\mathrm{FP}}\, \mathrm{fp}_{e,c} \right] \quad (1)$$

Here $fn_{e,c}$ and $fp_{e,c}$ are binary indicators of FN and FP errors for PCB c, and w^{FN} and w^{FP} control their relative importance. These weights are part of the evaluation, not the model.

To compare demographic groups, we compare case-adjusted disparity in mean encounter-level error. For groups a and a',

$$\Delta(a, a') = \sum_{v} \pi(v) \Big| \mathbb{E}[L(e) \mid A{=}a, V{=}v] - \mathbb{E}[L(e) \mid A{=}a', V{=}v] \Big| \quad (2)$$

where V is the vignette and $\pi(v)$ is the proportion of encounters in the full dataset that belong to vignette v, where $\sum_v \pi(v) = 1$.

4 Experiments

We evaluate ACTA with two views of fairness. First, we report PCB-level metrics by demographic group: recall, precision, F1, and positive prediction rate. Second, we compute the encounter-level disparity measure $\Delta(a, a')$ under multiple FN/FP weight ratios to test whether fairness conclusions are stable to different error priorities. Uncertainty is estimated with 1,000 bootstrap resamples at the encounter level, stratified by case so that case composition is preserved.

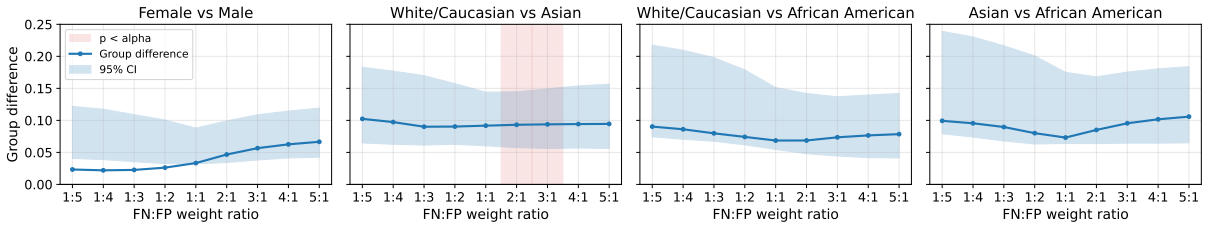

Figure 1: Absolute encounter-level disparity across different FN and FP weightings. The x-axis shows the FN/FP weight ratio, and the y-axis shows the absolute disparity estimate for each pairwise comparison.

Attribute	Group	Recall	Prec.	F1	PPR
Gender	Female	0.963	0.926	0.944	0.661
Gender	Male	0.960	0.908	0.933	0.633
Race	White/Caucasian	0.967	0.928	0.947	0.637
Race	Asian	0.931	0.915	0.923	0.637
Race	African American	0.979	0.906	0.941	0.677

Table 2: PCB-level performance metrics for the high-volume demographic groups in CLA.

Pair tested	$\Delta(a, a')$	95% CI	p-value
Female vs Male	0.016	[0.000, 0.108]	0.723
White/Caucasian vs Asian	0.090	[0.015, 0.182]	0.082
White/Caucasian vs African American	0.021	[0.000, 0.160]	0.748
Asian vs African American	0.055	[0.000, 0.205]	0.478

Table 3: Encounter-level disparity estimates under equal weighting of FNs and FPs. None of the pairwise comparisons were statistically significant.

Table 2 shows that the PCB-level metrics are broadly similar across the high-volume groups, with only small gender differences. Race shows somewhat more descriptive variation, especially in recall for comparisons involving Asian students, but none of the pairwise group differences is statistically significant after Bonferroni correction. Table 3 shows that the same general pattern holds for the encounter-level disparity measure under equal weighting of FNs and FPs: none of the pairwise comparisons is significant after correction, although the largest baseline disparity is observed for the White-Asian comparison.

Figure 1 shows that this baseline view is not fully stable across error-priority settings. As the relative weight assigned to FNs increases, absolute encounter-level disparity increases for all four pairwise comparisons. This indicates that group differences become larger when missed credit is treated as more costly than undeserved credit. At the same time, the comparisons do not behave identically: the White-Asian disparity is the only one that reaches significance anywhere in the sweep, doing so at the 2:1 and 3:1 FN:FP settings, while the remaining comparisons remain below the pairwise significance threshold. The White/Caucasian–Asian result seems to be mainly about false negatives. In Table 2, the Asian subgroup has lower recall than the White/Caucasian subgroup, while precision is more similar. This means the model misses more true PCBs for Asian learners, but it does not show an equally large difference in false positives. Because of that, the group difference becomes larger when false negatives are given more weight. We cannot say from this analysis why those missed detections happen. For this reason, we interpret the result cautiously as a local sensitivity to FN weighting, not as evidence of a broad or stable demographic disparity.

5 Discussion

The main finding from this study is that when errors are treated symmetrically—i.e., when FPs and FNs are assigned equal weight—the empirical results indicate no statistically significant differences across demographic groups. This suggests that, under a value framework in which both types of errors are considered equally consequential, the scoring system performs comparably across groups. This finding is particularly relevant because many intended uses of these scores assume such symmetry in error costs.

Next, we investigate patterns when weights on different types of errors are disproportionately increased. When increasing the weights on FPs up to fivefold, no statistically significant differences between groups are observed, suggesting that under a value framework that emphasizes high accuracy, the system performs equitably. A similar pattern holds when increasing the weights of FNs, with one exception: statistically significant differences emerge when FNs are assigned two- or threefold weights in the White–Asian comparison. We emphasize that we do not currently have a clear theoretical explanation for this pattern, and thus interpret it with caution, particularly in light of the fact that the effect disappears at higher weights (e.g., four- or fivefold). One possible account is that it

may be an example of Type I error resulting from multiple comparisons, despite the application of Bonferroni correction. It may also reflect a combination of sampling variability, subgroup imbalances, and underlying distributional differences in scores that render certain comparisons more sensitive to moderate changes in FN weighting. In particular, such reweighting may disproportionately affect borderline cases, producing localized effects. Overall, these findings point to a complex interaction between data characteristics and metric design, with a likely minimal practical effect.

6 Limitations and Ethical Considerations

Several limitations should be considered when interpreting the findings of this work.

First, the dataset is relatively limited in size, particularly for certain subgroup comparisons. This constrains statistical power and reduces the ability to detect small but potentially meaningful differences across demographic groups. This limitation is especially relevant for intersectional analyses, where subgroup counts become even smaller.

Second, the analysis is conducted within the specific context of the ACTA scoring system and the CLA learning tool. While these provide a valuable and realistic testbed, they represent a particular type of automated scoring application with its own design choices and use cases. Consequently, the findings may not generalize to other automated scoring settings, such as those involving different constructs, item formats, model architectures, or deployment contexts. Replication across diverse systems and domains is necessary to establish the broader applicability of these results.

Third, the weighting schemes used in this study are intentionally simple, focusing on systematically varying the relative importance of FPs and FNs. While this approach is useful for conducting sensitivity analyses and illustrating how fairness conclusions shift under different value assumptions, it does not fully capture the complexity of real-world assessment policies. In practice, decisions about error costs may depend on a wider range of factors, including the stakes of the assessment, the distribution of scores, downstream decision-making processes, and stakeholder priorities. More nuanced or context-specific cost functions could lead to different patterns of results.

In terms of ethical considerations, institutions deploying automated scoring systems have a responsibility to examine disaggregated error patterns, assess the potential harms associated with different error types, and ensure that mechanisms exist for identifying and correcting erroneous scores. Such analyses help ensure that automated scoring systems do not introduce or amplify inequities through the unequal distribution of scoring errors.

A key consideration is that while most investigations of group differences treat FP and FN errors as equally important, the intended use of an assessment should guide the relative weight of these errors. The consequences of awarding credit without merit (FP) or withholding credit when it's deserved (FN) can differ substantially in low vs. high stakes and formative vs. summative assessments. For example, in a high-stakes context like a medical licensure exam, incorrectly passing an unqualified physician (FP) can potentially result in risk to real patients. In this case, minimizing FP errors might be prioritized. In contrast, for low-stakes formative tools, such as CLA, failing to award credit when it's deserved (FN) can affect a learner's confidence and motivation, ultimately undermining the main purpose of the learning tool.

These differences underscore that error weighting is fundamentally a value-laden decision, tied to the goals and consequences of the assessment. Treating FP and FN errors symmetrically may be appropriate in some cases, but it should not be assumed as a default. Instead, evaluation frameworks should explicitly align error trade-offs with the intended use of the assessment, as different choices can lead to materially different conclusions about system performance and fairness across groups.

A remaining source of uncertainty is that both human annotation and automated scoring operate on the same de-identified text. As a result, construct-irrelevant linguistic variation correlated with demographic group could affect both human labels and model predictions. Potential sources include subgroup differences in lexical, grammatical, or discourse patterns, as well as uneven subgroup representation in model training.

References

Nico Andersen, Julia Mang, Frank Goldhammer, and Fabian Zehner. 2025. Algorithmic fairness in automatic short answer scoring. *International Journal of Artificial Intelligence in Education*, pages 1–38.

Indrani Dey, Dana Gnesdilow, Rebecca Passonneau, and Sadhana Puntambekar. 2024. Potential pitfalls

of false positives. In *International Conference on Artificial Intelligence in Education*, pages 469–476. Springer.

Helen M Ernst, Anja Prinz-Weiß, Jörg Wittwer, and Thamar Voss. 2025. Discrepancy between performance and feedback affects mathematics student teachers' self-efficacy but not their self-assessment accuracy. *Frontiers in Psychology*, 15:1391093.

Gerardo Flores, Alyssa Hasegawa Smith, Julia Fukuyama, and Ashia C. Wilson. 2025. Aligning evaluation with clinical priorities: Calibration, label shift, and error costs. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2021. The (im) possibility of fairness: Different value systems require different mechanisms for fair decision making. *Communications of the ACM*, 64(4):136–143.

Matthew S Johnson, Xiang Liu, and Daniel F McCaffrey. 2022. Psychometric methods to evaluate measurement and algorithmic bias in automated scoring. *Journal of Educational Measurement*, 59(3):338–361.

Matthew S Johnson and Daniel F McCaffrey. 2023. Evaluating fairness of automated scoring in educational measurement. *Advancing natural language processing in educational assessment*, 142.

Tiffany Wenting Li, Silas Hsu, Max Fowler, Zhilin Zhang, Craig Zilles, and Karrie Karahalios. 2023. Am I wrong, or is the Autograder wrong? Effects of AI grading mistakes on learning. In *Proceedings of the 2023 ACM Conference on International Computing Education Research-Volume 1*, pages 159–176.

Diane Litman, Haoran Zhang, Richard Correnti, Lindsay Clare Matsumura, and Elaine Wang. 2021. A fairness evaluation of automated methods for scoring text evidence usage in writing. In *International Conference on Artificial Intelligence in Education*, pages 255–267. Springer.

Anastassia Loukina, Nitin Madnani, and Klaus Zechner. 2019. The many dimensions of algorithmic fairness in educational applications. In *Proceedings of the fourteenth workshop on innovative use of NLP for building educational applications*, pages 1–10.

Saed Rezayi, LA Ha, Y Zhou, et al. 2025. Automated scoring of communication skills in physician–patient interaction: Balancing performance and scalability. In *Proc 20th BEA Workshop*.

Nils-Jonathan Schaller, Yuning Ding, Andrea Horbach, Jennifer Meyer, and Thorben Jansen. 2024. Fairness in automated essay scoring: A comparative analysis of algorithms on german learner essays from secondary education. In *Proceedings of the 19th workshop on innovative use of nlp for building educational applications (bea 2024)*, pages 210–221.

A Annotation Guideline

1. Annotate Complete Components. Ensure each annotation captures the complete text relevant to the label, even if the sentence is grammatically incomplete.

2. Annotate all instances of a concept or label. More variations of a label or concept will better train the automated scoring model.

3. Include Filler Words Within a Text Span. Stutters, repetitions, false starts, and filler words such as 'um,' 'ok,' or 'like' should be included only if they are in the middle of a text span. Do not annotate them if they are at the beginning or end.

4. Do Not Infer Beyond the Text. Annotate only what is explicitly stated. Avoid making assumptions or interpretations beyond the text.

5. Maintain Neutrality. Do not let personal opinions or interpretations influence the annotation. Focus solely on the content and structure. Transcripts are not being scored; they serve as a source of information to train an automated scoring engine.

6. Follow Label Definitions Strictly. Always refer to the label definitions provided. If a segment doesn't clearly fit a label, leave it unannotated or flag it for review.

7. Parsimony. Annotate only the minimal number of words necessary to accurately represent the label, avoiding extraneous or redundant text to maintain clarity and consistency across annotations.

8. Separate Labels for Adjacent Components. When two adjacent statements or questions both correspond to the same label, annotate each one separately.

9. Punctuation Does Not Affect Annotation. Include or exclude it as needed, focusing instead on accurately capturing the relevant words for the label.

10. Avoid Overlap. Annotations can sometimes seem to have similar concepts. However, if there is overlap in the annotation, double-check the label definition and/or flag for review, as there should be a distinction and minimal overlap.

case id	174	175	176	177	178	180	182	192
n encounters	54	52	54	59	46	55	58	59
n PCBs	3	1	3	4	3	3	4	4
	Number of positive PCB annotations (prevalence rate in %)							
By Gender								
Male	91 (56%)	52 (100%)	80 (49%)	164 (69%)	56 (41%)	99 (60%)	171 (74%)	134 (57%)
Female	48 (52%)	33 (100%)	41 (46%)	98 (72%)	32 (41%)	63 (64%)	109 (72%)	90 (58%)
By Ethnicity								
White/Caucasian	55 (59%)	31 (100%)	39 (45%)	98 (70%)	34 (42%)	54 (55%)	97 (73%)	78 (57%)
Asian	21 (54%)	11 (100%)	27 (56%)	32 (73%)	15 (38%)	22 (67%)	34 (71%)	31 (52%)
African American	8 (44%)	6 (100%)	10 (56%)	26 (72%)	2 (33%)	12 (67%)	24 (75%)	7 (44%)

Table 4: Number of encounters, and PCBs, and positive annotations by case, gender, and ethnicity.

B Data Statistics

Table 4 summarizes the eight cases included in the analysis. The number of encounters is similar across cases, ranging from 46 to 59, while the number of PCBs ranges from 1 to 4. Positive annotation rates vary across cases, but differences by gender and ethnicity within each case are generally small.