

DOI: https://doi.org/10.48009/3_iis_2026_140

Bias taxonomy and mitigation pipelines for AI-based educational assessment systems

Parvinder Pal Singh, *The University of Texas Permian Basin*, singh_p@utpb.edu

Carl M. Rebman Jr., *University of San Diego*, carlr@sandiego.edu

Mark McMurtrey, *University of Central Arkansas*, markmc@uca.edu

Abstract

Algorithmic assessment systems now score essays, generate feedback, monitor remote examinations, and classify academic risk. These systems offer rapid processing, consistency, and scalability, but that same efficiency carries a technical risk, i.e., a biased model can reproduce the same error pattern consistently, quietly, and at institutional scale, in ways that are difficult to detect. This paper develops the Bias Taxonomy and Mitigation Pipeline (BTMP) framework, a fairness-engineering framework for AI-based educational assessment systems. The framework is for developers, vendors, and technical reviewers who build, tune, procure, and audit these systems. It distinguishes data-related, algorithmic, and interaction-driven sources of bias using a taxonomy built with a stated development method; maps fairness criteria to specific assessment contexts, differentiated by decision stakes as well as by task; and organizes mitigation strategies into pre-processing, in-processing, and post-processing stages. Its central contributions are the context-to-harm-to-metric mapping, a consolidated diagnostic and testing procedure, risk-based deployment tiers, and a minimum evidence package for technical approval. The paper connects this taxonomy to the educational-measurement tradition alongside the machine-learning fairness literature and illustrates its diagnostic procedure with a hypothetical worked example on a public essay-scoring benchmark.

Keywords: algorithmic fairness, educational measurement, automated assessment, bias mitigation, fairness metrics, machine learning, taxonomy development

Introduction

Machine scoring and monitoring systems now make judgments that were once handled almost entirely by instructors, graders, and academic integrity officers. In educational assessment, artificial intelligence is used to score essays, classify risk, recommend interventions, generate formative feedback, and flag suspicious behavior during remote examinations. These applications are not hypothetical. The e-rater engine has been used operationally, alongside human raters in a check-score process, to score the Analytical Writing measure of the GRE General Test since 2008 (Attali & Burstein, 2006; Breyer et al., 2014). Automated writing evaluation tools such as MI Write are deployed at scale in elementary and secondary classrooms, including with English language learners (Wilson & Huang, 2024). Remote proctoring tools saw a documented surge in adoption during the COVID-19 pandemic, and a substantial share of instructors who adopted them in 2020-21 report intending to continue using them after a return to in-person instruction (Balash et al., 2023). These deployments carry real and asymmetric stakes. A false positive in proctoring

can trigger an academic misconduct investigation, while a false negative in a learning analytics system can cause a struggling student to be overlooked.

The challenge is not that AI systems may be imperfect but that machine learning systems can transform local assessment irregularities into systematic subgroup disparities. Human assessment has always had its own biases, inconsistencies, and judgment issues; this paper does not claim automation introduces bias where none existed before. It frames educational AI fairness as a technical and socio-technical design problem. Prior fairness work has produced important mathematical criteria, including equalized odds, predictive parity, individual fairness, and counterfactual fairness, but these criteria are often discussed apart from the assessment settings in which they would be applied, and apart from the older measurement fairness tradition that assessment already has. Automated essay scoring, proctoring, formative feedback, and predictive analytics produce different kinds of harm and therefore require different fairness tests, treating these applications as equivalent weakens fairness evaluation.

This paper is written as fairness-engineering guidance for the people who build, tune, procure, and technically audit AI assessment systems, i.e., developers, vendors, and technically informed reviewers. It specifies where in the build-develop-validate-deploy-maintain lifecycle a technical reviewer intervenes and what evidence that intervention should produce. The paper introduces the Bias Taxonomy and Mitigation Pipeline (BTMP) framework, which has four functions. First, it classifies the bias mechanisms most relevant to AI-based educational assessment, using a taxonomy built with a stated development method. Second, it aligns fairness metrics with the decision context, the decision's stakes, and the likely student harm. Third, it organizes mitigation techniques into pre-processing, in-processing, and post-processing stages so that intervention points are visible across the model lifecycle. Fourth, it specifies the evidentiary basis on which a technical approval decision can be made.

Background and Technical Foundation

Algorithmic bias refers to systematic differences in model outputs or error rates that can disadvantage particular groups or individuals. In educational assessment, bias may appear as lower scoring accuracy for one linguistic group, higher false-flag rates for one group of test takers, or feedback that varies in quality based on inferred student characteristics. The broader machine learning literature emphasizes that bias can emerge from training data, labeling practices, feature choices, optimization objectives, and deployment feedback loops (Mehrabi et al., 2021).

Automated essay scoring illustrates the problem clearly. Early systems relied heavily on surface linguistic features such as syntax, grammar, sentence length, and lexical complexity (Attali & Burstein, 2006). These features may correlate with writing quality in many datasets, but they can also reward conformity to dominant academic language norms. If the system treats linguistic difference as error, it may be measuring standard language alignment rather than writing ability. Recent research on automated essay scoring and writing feedback has continued to identify risks associated with linguistic diversity, demographic inference, and uneven feedback quality (Fan & Yun, 2026; Tan et al., 2026; Wilson & Huang, 2024). Wilson and Huang's (2024) study of automated versus human scoring for elementary age English language learners is especially relevant to the diagnostic procedure. It found that automated scores were less predictive of external writing test outcomes for English language learners than for other students, but that human raters showed the same weaker relationship, indicating the automated system was not uniquely biased but was reproducing a bias already present in its human-generated training labels. Similarly, Tan, Phalen, and Demszky (2026) find that large language models used to generate personalized writing feedback shift the content of that feedback, in stereotype aligned ways, based on a student's presumed race, gender, and other attributes, even when the underlying essay is identical.

Remote proctoring systems introduce a different technical risk. Proctoring tools often depend on facial detection, gaze estimation, identity verification, and behavioral classification. A direct study of automated proctoring software documented statistically significant racial, skin-tone, and sex disparities in flagging behavior (Yoder-Himes et al., 2022), a pattern consistent with broader evidence that commercial facial-analysis systems perform unevenly across intersecting demographic groups when training data are not representative (Buolamwini & Gebru, 2018). In an educational setting, unequal technical performance can become unequal institutional exposure to suspicion, a model that fails more frequently for some students can create higher interruption, escalation, or false accusation risk for those students specifically.

Prior Work

The nearest prior work to this paper's premise is Loukina, Madnani, and Zechner (2019), who lay out multiple dimensions of algorithmic fairness for educational applications and reason about which fairness definitions suit which testing use cases. Baker and Hawn (2022) provide the most comprehensive empirical review of documented algorithmic bias in education, cataloguing which student subgroups are affected and at which stage of development and deployment bias has been observed. Suresh and Gutttag (2021) offer a general machine-learning lifecycle framework identifying seven sources of downstream harm, spanning data collection through deployment, that is not specific to education.

Prior Work	Focus	What This Paper Adds
Loukina, Madnani, & Zechner (2019)	Surveys fairness dimensions and reasons about metric choice for educational NLP applications.	Operationalizes these dimensions into a stakes-differentiated context-to-metric mapping (Table 3), a diagnostic procedure (Table 5), and an evidence package for technical approval (Table 7).
Baker & Hawn (2022)	Empirical review of documented algorithmic bias across education, organized by affected subgroup and pipeline stage.	Adds a bias taxonomy with a stated development method (Table 2) and a mitigation pipeline organized explicitly by lifecycle stage (Table 4; Figure 2).
Suresh & Gutttag (2021)	General, non-education seven-source machine-learning harm framework spanning the ML lifecycle.	Maps a domain-specific three-family taxonomy and mitigation pipeline onto this lifecycle for educational assessment specifically.
AERA/APA/NCME Standards (2014); Messick (1989)	Measurement-fairness apparatus: differential item functioning, measurement invariance, construct-irrelevant variance, predictive vs. measurement bias.	Connects this apparatus directly to ML-fairness metrics (Table 3) and treats it as a required part of technical fairness review, not an optional add-on.

Table 1. Positioning Relative to Prior Fairness-in-Education Work

This paper's contribution sits alongside, not in place of, these three. It takes Loukina et al.'s fairness dimensions reasoning and Suresh and Gutttag's lifecycle orientation and operationalizes them for AI-based educational assessment specifically. Table 1 summarizes this positioning where a stakes-differentiated context-to-metric mapping, a taxonomy built with a stated development method, a consolidated diagnostic and testing procedure, and a minimum evidence package tied to a specific approval decision are this paper's technical contribution.

The Educational-Measurement Tradition

Because this is a paper about assessment, bias should not be treated only in the generic machine-learning sense. Educational and psychological measurement has its own, older fairness apparatus. The Standards for Educational and Psychological Testing, jointly published by AERA, APA, and NCME (2014), distinguish ‘predictive bias’ which is a test that predicts an external outcome differently across groups from ‘measurement bias’ which is a test that yields systematically different scores for equally able members of different groups. The diagnostic procedure in Table 5 uses this distinction directly, where subgroup performance evaluation (Step 4) targets predictive bias, while label review (Step 3) and construct clarification (Step 1) target measurement bias. The Standards also formalize differential item functioning (DIF), a family of statistical techniques for detecting when an individual item behaves differently for examinees of equal ability from different groups, and measurement invariance which is the broader property that a test measures the same construct in the same way across groups. Both techniques predate machine learning fairness metrics by decades and apply directly to automated scoring which is a scoring feature that functions like an item can be tested for DIF using matched examinee ability, and a scoring model can be tested for measurement invariance using approaches already applied to human-scored rubrics.

Construct-irrelevant variance, a term this paper uses throughout, is Messick’s (1989) language for one specific threat to validity which is score variance driven by features unrelated to the intended construct. Its companion threat, construct underrepresentation, occurs when a model captures only part of the intended construct, for example, a model that scores grammar and organization well but cannot recognize argument quality underrepresents the writing construct even if no subgroup shows disparate error. Engaging this tradition changes what fairness means for an assessment system. A system can satisfy every machine-learning fairness metric in Table 3 and still be invalid if it is not measuring the construct it claims to measure. The remainder of this paper treats the measurement tradition and the machine-learning fairness tradition as jointly necessary, not as alternatives.

Fairness Criteria and Tradeoffs

Fairness criteria are not interchangeable. Chouldechova (2017) and Kleinberg et al. (2017) show that several desirable fairness conditions cannot generally be satisfied at the same time when base rates differ across groups. This impossibility result matters for education because student outcome distributions are shaped by unequal prior opportunity, access, and preparation; a technical team cannot simply declare a model fair by selecting one convenient measure. Equal opportunity is most relevant when the central concern is whether qualified students are recognized at equal rates (Hardt et al., 2016). Equalized odds is more demanding because it considers both false positive and false negative rates across groups. Predictive parity emphasizes whether a positive score or flag has the same meaning across groups. Individual fairness asks whether similar students receive similar outputs (Dwork et al., 2012), while counterfactual fairness examines whether an output would change if a protected attribute were altered while relevant features remained constant (Kusner et al., 2017). In educational assessment, these metrics should be treated as a menu of context-specific diagnostics rather than as a single universal standard.

Most of the classification based criteria above including equal opportunity, equalized odds, predictive parity, and predictive equality presuppose a discrete decision such as “qualified” or “flagged.” Proctoring flags are a natural fit because the underlying decision is already binary. Essay scores are usually ordinal or continuous, so applying these criteria requires either thresholding a score into a pass/fail-style decision, which discards information and can itself introduce bias at the threshold, or substituting fairness notions defined for ordinal and continuous outcomes, such as subgroup differences in mean scoring error, subgroup calibration curves, or rank based parity measures. Table 3 shows the outcome type each context assumes, where a table entry recommends a classification style metric for an ordinal or continuous context, that recommendation should be read as requiring an explicit thresholding decision or an ordinal metric substitute, not as a claim that the metric applies unmodified.

Bias Taxonomy for AI-Based Educational Assessment

The taxonomy below was developed following the iterative method for taxonomy development in information systems proposed by Nickerson, Varshney, and Muntermann (2013), which combines a conceptual-to-empirical pass, deriving candidate dimensions from the fairness and educational measurement literatures with an empirical-to-conceptual pass that checks candidate dimensions against documented cases of bias in automated essay scoring and proctoring systems. The meta-characteristic used to group bias mechanisms is the stage of the system lifecycle at which the mechanism is introduced, i.e., before training (data-related), during model development (algorithmic), or after deployment, through the system's interaction with students and instructors (interaction).

Applying this meta characteristic consistently resolves an ambiguity present in an earlier version of this taxonomy, in which confirmation bias, a human cognitive tendency to favor information that supports a prior belief had been grouped with algorithmic bias despite being a property of human judgment rather than of the model, and in which feedback loop effects had appeared in two families at once. Confirmation bias is placed here in the interaction family, since it describes how instructors and students respond to model outputs after deployment, not how the model is built. Feedback loop effects appear only in the interaction family, where a feedback loop originates in a modeling choice, such as an objective function retrained on its own prior outputs, that mechanism is classified as optimization bias in the algorithmic family, and only its downstream, human mediated consequences are classified as interaction bias. Each subtype in Table 2 is intended to belong to exactly one family under this rule.

Bias Family	Subtypes	Technical Concern
Data-Related Bias	Historical bias, representation bias, measurement bias	Training examples, labels, and proxies encode inequality before model training.
Algorithmic Bias	Design bias, optimization bias	Feature choices, objectives, and model updates produce unequal subgroup behavior.
Interaction Bias	Feedback loops, confirmation bias (human over trust of model outputs), stereotyping / exclusion effects	Human use, student adaptation, and access conditions change how model outputs affect learners after deployment.

Table 2. Bias Taxonomy for AI-Based Educational Assessment Systems (BTMP Framework)

Data-related bias arises before model training begins. Historical bias occurs when training records reflect prior inequality rather than neutral distributions of ability. Representation bias occurs when groups are insufficiently represented in training data, producing weaker generalization for those groups. Measurement bias occurs when labels or proxies do not measure the intended construct equivalently across students; in assessment, this is especially important because scores are supposed to represent ability or performance, not conformity to irrelevant features.

Algorithmic bias is introduced through modeling choices. Feature selection may reward attributes that are easy to measure but only weakly connected to the construct being assessed. Optimization bias occurs when a model is tuned to maximize aggregate accuracy while subgroup disparities remain hidden. These risks are not usually the result of explicit discriminatory intent, they are often byproducts of ordinary model development practices.

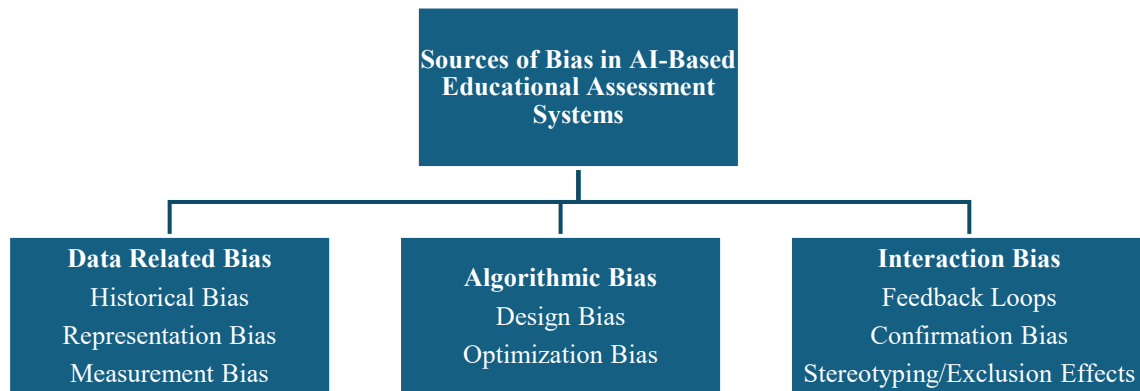


Figure 1. Bias Taxonomy for AI-Based Educational Assessment Systems

Interaction bias emerges after deployment. Instructors may overtrust a score because it appears objective, a form of automation bias that functions, in this taxonomy, as the interaction family analogue of confirmation bias. Instructors and students give model outputs more evidentiary weight than the model's own validation evidence supports. Students may learn to write for the scoring engine rather than for meaningful communication. Proctoring systems may assume access to quiet rooms, reliable internet, quality cameras, and stable lighting; students who lack these conditions can be excluded or flagged for reasons unrelated to academic integrity. These interaction effects are not solved by model retraining alone, because the bias exists in the connection between the system, the user, and the educational environment, addressing it requires interface design, instructor training, and disclosure alongside any technical mitigation.

Context-to-Harm-to-Metric Mapping

The framework links fairness metrics to the type of educational decision being automated and to that decision's stakes, since the same task can carry very different consequences depending on where it sits in a student's academic trajectory. A K-12 formative writing tool that only the student sees carries a different tolerable error rate than a licensure or admissions essay that determines who is credentialed. A low stakes practice quiz proctor carries different consequences than a certification exam proctor. Table 3 therefore differentiates automated essay scoring and AI proctoring by stakes, in addition to distinguishing them from predictive analytics and automated feedback, and pairs each context with its primary harm, its recommended metric, that metric's technical definition, and the underlying outcome.

Application / Context	Primary Fairness Harm	Recommended Metric(s)	Technical Definition	Outcome-Type
K-12 / formative essay scoring (student-facing practice tool, no grade attached)	Unequal recognition of writing competence in low-stakes feedback	Equal opportunity	Similar true positive rates among qualified students	Ordinal/continuous (thresholding required to compute rates)

Application / Context	Primary Fairness Harm	Recommended Metric(s)	Technical Definition	Outcome-Type
High-stakes essay scoring (grading, admissions, or licensure)	Unequal recognition of competence with consequential impact; unequal score meaning	Equal opportunity; predictive parity	Equal opportunity as above. Predictive parity: a positive score has similar meaning across groups	Ordinal/continuous (requires thresholding or an ordinal-metric substitute (e.g., subgroup mean error, calibration curves))
Low-stakes AI proctoring (practice exams, ungraded quizzes)	False accusation risk, though consequences are limited	Equalized odds	Similar true positive and false positive rates across groups	Binary (flag / no flag) - classification metrics apply directly
High-stakes AI proctoring (certification, licensure, high-stakes exams)	False accusation or unequal investigation burden with severe consequences	Equalized odds; predictive equality	Equalized odds as above. Predictive equality: similar false positive rates across groups	Binary (classification metrics apply directly)
Predictive analytics (risk flags, advising)	Unequal access to intervention or opportunity, compounding over time	Counterfactual fairness; individual fairness	Counterfactual: output unchanged if a protected attribute changed with relevant traits held constant. Individual: similar students receive similar outputs	Typically, binary or ordinal risk tier
Automated feedback (personalized, LLM-generated)	Differential quality, tone, or substance of feedback	Subgroup content audits; counterfactual testing	No single closed-form metric; requires paired-prompt testing of feedback content across presumed student attributes (see Tan et al., 2026)	Free text output (requires content analysis, not a rate-based metric)

Table 3. Context-to-Harm-to-Metric Mapping for Educational AI Assessment

This mapping also shows why metric pluralism is necessary. Statistical parity is attractive because it is simple, but it does not necessarily reflect educational validity when base rates differ for reasons unrelated to the model. Equal opportunity fits a scoring system but does not fully address false-accusation risk in proctoring. Predictive parity helps interpret scores but can conflict with equalized odds when base rates differ across groups. Technical teams should therefore document not only the metric selected but also the metrics not selected and the reason for excluding them. High stakes systems should report at least two complementary diagnostics. A proctoring system might report equalized odds and predictive equality, an essay scoring system might report equal opportunity and predictive parity, and a learning analytics system might combine counterfactual fairness with individual fairness. Fairness thresholds should be stated in advance, i.e., what subgroup performance difference is acceptable, what difference requires mitigation, and what difference requires suspension or redesign, since without stated thresholds, fairness reporting becomes descriptive rather than actionable.

Bias Mitigation Pipeline

The organization of mitigation into pre-, in-, and post-processing stages is standard in the algorithmic fairness literature (Agarwal et al., 2018; Kamiran & Calders, 2012; Zhang et al., 2018) rather than a contribution unique to this paper. What this paper adds is the systematic mapping of that standard taxonomy onto specific educational assessment contexts, stakes, and lifecycle control points (Tables 3, 5, and 6), and the evidence package that operationalizes it for technical approval (Table 7). A serious fairness evaluation should still consider all three stages rather than treating mitigation as a single correction applied at the end.

Pre-processing methods change the data before training and are appropriate when the dataset is the main source of the fairness problem. Reweighting increases the influence of underrepresented cases during training. Resampling changes group distributions by oversampling minority cases or undersampling majority cases. Data augmentation can create additional examples for underrepresented groups, although generated examples must be validated to avoid introducing artificial patterns. Kamiran and Calders (2012) demonstrate that pre-processing can reduce discrimination in classification.

Stage	Examples	Primary Use
Pre-processing	Reweighting, resampling, data augmentation	Best when the problem is dataset imbalance or underrepresentation.
In-processing	Fairness regularization, adversarial debiasing, representation learning	Best when model objectives or latent representations encode protected information.
Post-processing	Threshold adjustment, calibration, human review	Best when model access is limited or institutional decision rules must be adjusted.

Table 4. Three-Stage Bias Mitigation Pipeline

In-processing approaches intervene during model training. Fairness regularization adds a penalty to the objective function so the model is discouraged from relying on representations associated with sensitive attributes. Reductions based approaches formalize fairness constraints as optimization problems (Agarwal et al., 2018). Adversarial debiasing trains a primary model to perform the assessment task while an adversary attempts to infer group membership and the primary model is penalized when group membership remains recoverable (Zhang et al., 2018). Representation learning approaches similarly seek latent spaces in which task relevant information is retained while protected information is minimized (Zemel et al., 2013). In automated essay scoring specifically, Fan and Yun (2026) apply this approach to ESL grading bias. After documenting that a fine-tuned scoring model penalized high proficiency ESL essays relative to native speaker essays of identical human rated quality, they fine tune the model using a triplet-margin loss over matched essay pairs so that the model's latent representations cluster essays by writing quality rather than by linguistic origin. This is an in-processing intervention rather than a pre-processing data-construction technique, since the mitigation acts on the model's learned representations during training rather than on the training data itself.

Post-processing methods are often the most accessible to institutions using vendor systems because they operate on scores, flags, or thresholds without requiring model access. Threshold adjustment can reduce unequal false positive or false negative rates across groups (Hardt et al., 2016). Calibration can align the meaning of output scores, although calibration may conflict with other fairness criteria when base rates differ (Pleiss et al., 2017). Human review is not a purely technical correction, but it functions as a post-processing safeguard when automated outputs would otherwise trigger high stakes decisions on their own.

Fairness Diagnostic and Testing Procedure

A fairness framework is only useful if paired with a procedure that can be followed consistently and audited afterward. The eight steps in Table 5 consolidate what earlier versions of this framework presented separately as a diagnostic procedure and a testing protocol. Each step below folds together the diagnostic question a technical reviewer must answer and the technical test that answers it.

Step	Diagnostic Question	Technical Test
1. Construct clarification	What educational construct is this system meant to measure, and what technical proxy stands in for it?	Baseline check: does the model align with the stated construct, not merely with historical scores?
2. Dataset accounting	Do the intended population, training population, and validation population match?	Documentation review: source, time period, task type, language, demographic representation, and known gaps.
3. Label review	Do training labels (human scores, misconduct flags) already encode bias?	Subgroup inter-rater reliability compared against subgroup model error (Wilson & Huang, 2024).
4. Subgroup performance evaluation	Does performance differ by relevant student characteristics or conditions?	Subgroup performance test with reported sample sizes and confidence intervals; no-disparity claims are not made when the design lacks power to detect one. Evaluate major intersectional groups when statistical power permits.
5. Sensitivity testing and harm mapping	Do irrelevant changes (dialect, lighting, camera quality) alter outputs, and what is the consequence of each likely error?	Sensitivity test using matched cases that hold the construct constant while varying nonconstruct features; metric selection justified against the mapped harm (Table 3).
6. Mitigation selection	Which stage - pre-, in-, or post-processing - matches the diagnosed source of bias?	Documented mitigation history with a before/after subgroup comparison (Table 4).
7. Deployment simulation	How does the system behave under realistic conditions, including negative cases the vendor does not showcase?	Deployment simulation using negative cases (e.g., strong reasoning with nonstandard syntax; low light; assistive technology).
8. Monitoring and retesting	When must the system be re-evaluated?	Retesting triggers: model updates, expansion to new courses or populations, a change in stakes, or a rise in complaints.

Table 5. Fairness Diagnostic and Testing Procedure

Two points from this procedure deserve emphasis because they are easy to skip under deadline pressure. First, negative-case testing (Step 7) is not optional. Developers routinely demonstrate a system working under ideal conditions, but institutions need to know when it fails, and vendor demonstrations rarely disclose this. Second, fairness testing must be repeatable, not a one-time vendor demonstration, because models, student populations, prompts, and deployment conditions all change, Step 8 specifies the events that should trigger a repeat of Steps 3 through 5 rather than leaving re-evaluation to institutional discretion.

Application to Automated Essay Scoring

Automated essay scoring is a useful test case because it combines measurement, language, and model design issues, and because it spans the full stakes range in Table 3, from ungraded formative practice to admissions and licensure decisions. A system may be trained to predict human scores, but the human scores may already reflect assumptions about what counts as effective writing, the model may then learn surface features that correlate with those scores. This does not make the system useless, but it means the system must be examined for construct alignment. If the intended construct is argument quality, a model that rewards sentence length or standard grammar more strongly than reasoning creates measurement bias in the sense defined by the Standards (AERA, APA, & NCME, 2014). A technical review should begin with feature interpretation. Developers should document whether the model relies on lexical sophistication, syntax, grammar errors, discourse structure, evidence use, semantic similarity, or prompt-specific patterns. For opaque models, reviewers should still require sensitivity testing, using matched essays that hold argument structure constant while varying language background or dialect features, if scores change in ways unrelated to the intended construct, the model is vulnerable to construct-irrelevant variance.

Automated essay scoring also requires careful label review. If human raters are inconsistent across groups or prompts, model agreement with human scores does not establish fairness. High agreement can simply mean the model has learned the same biased scoring pattern, as Wilson and Huang (2024) document directly. For this reason, subgroup interrater reliability and subgroup model error should both be examined. Mitigation may begin with pre-processing if the training corpus underrepresents particular language backgrounds or writing styles, move to in-processing such as the representation learning approach of Fan and Yun (2026) when the model's latent representations themselves encode linguistic origin information, or use post-processing calibration to align score distributions, applied cautiously since score adjustments alone may not address an underlying construct problem.

The strongest technical evidence for a scoring system is not a single accuracy statistic but a package of documentation showing that the model performs adequately across groups and stakes tiers, supports the intended construct, and has been tested against plausible sources of linguistic bias using the procedure in Table 5.

Application to AI-Powered Proctoring

AI-powered proctoring differs from essay scoring because the central harm is exposure to suspicion rather than a missed learning opportunity, and because, as Table 3 reflects, the underlying decision is already binary, so classification-based fairness metrics apply without a thresholding step. The system may monitor face visibility, gaze direction, background movement, sound, browser activity, or identity verification, and these signals are transformed into flags or risk scores. Yoder-Himes et al. (2022) document that this transformation is not demographically neutral in deployed systems. Their study of commercial proctoring software found statistically significant racial, skin-tone, and sex disparities in flagging. From a technical fairness perspective, the central question is whether the system creates unequal false positives or unequal burden across groups and environments, and this question has a different answer for a low stakes practice quiz than for a certification or licensure exam.

Proctoring systems are particularly sensitive to deployment conditions. Lighting, camera quality, internet stability, room privacy, and disability related movement can all affect system behavior, and a tool that performs well under standardized test-lab conditions may perform poorly in the varied environments in which students actually take exams. Technical review should include environmental robustness testing, not only demographic subgroup testing. A fairness audit that ignores access conditions will miss exclusion bias. Equalized odds and predictive equality are the relevant diagnostics for high stakes proctoring, since false positives generate investigations, anxiety, and reputational harm. Mitigation for proctoring will often

involve post-processing and governance controls because institutions may not have access to the vendor model. Threshold adjustment can reduce unequal false flag rates, but it should not function as a hidden correction, and human review should be mandatory before any flag becomes an academic integrity allegation, with the reviewer given access to the evidence, the system's documented limitations, and the student's response.

Proctoring also requires continuous monitoring, since vendor updates can change flagging behavior and changes in student hardware or course modality can alter error patterns. Institutions should require periodic reports on false flag rates, appeal outcomes, technical failures, and accessibility issues. A proctoring system should be considered technically fair only if it performs reliably across the actual conditions under which students are expected to use it, not only under vendor demonstrated conditions.

Risk-Based Deployment Tiers

Not every AI assessment use requires the same level of technical scrutiny. A formative writing suggestion tool used only by students for practice creates less institutional risk than an automated scoring system that submits grades directly. A proctoring system that flags students for human review creates less risk than one that automatically blocks students from completing an exam. Table 6 supports risk-based deployment tiers that cross the stakes distinctions already introduced in Table 3. Low-risk uses include optional practice feedback, student facing revision suggestions, and instructor tools that are clearly advisory; these still require disclosure and basic validation but may not require formal subgroup auditing before every use. Medium-risk uses include AI tools that influence grading recommendations, advising priorities, or instructor attention, and require stronger documentation, faculty training, and monitoring because they shape human judgment. High-risk uses include systems that directly assign grades, initiate misconduct processes, or restrict academic opportunities, and require formal fairness evaluation, human review, appeal mechanisms, and ongoing audits and, where the minimum evidence package (Table 7) cannot be produced, withholding of technical approval until it can be, rather than deployment under a promise of retrospective evaluation.

Risk Tier	Example Use	Minimum Control
Low	Optional practice feedback or student self-review	Disclosure and basic validation
Medium	Instructor-facing recommendations or preliminary scores	Documentation, faculty training, subgroup checks
High	Grades, misconduct flags, progression, or access decisions	Formal fairness audit, human review, appeal process, continuous monitoring; withhold approval if the evidence package (Table 7) is incomplete

Table 6. Risk-Based AI Assessment Deployment Tiers

Risk tiering prevents two opposite errors. Overregulation of low-stakes instructional experimentation, which can discourage useful innovation, and under-review of high-stakes automation, which can expose students to hidden and consequential errors. A proportional framework lets institutions and developers focus technical review where the consequences of unfairness are greatest.

Lifecycle Controls, Boundary Conditions, and Vendor Opacity

The mitigation pipeline (Table 4) and diagnostic procedure (Table 5) operate across the system lifecycle rather than at a single point. During design, the control point is construct definition and dataset planning; during development, it is model training, fairness testing, and mitigation; during validation, it is subgroup evaluation and documentation; during deployment, it is human review, threshold policy, and monitoring; during maintenance, it is re-evaluation after updates, population shifts, or new uses. Figure 2 maps the three mitigation stages onto this lifecycle and shows the monitoring loop that re-enters development and validation after deployment.

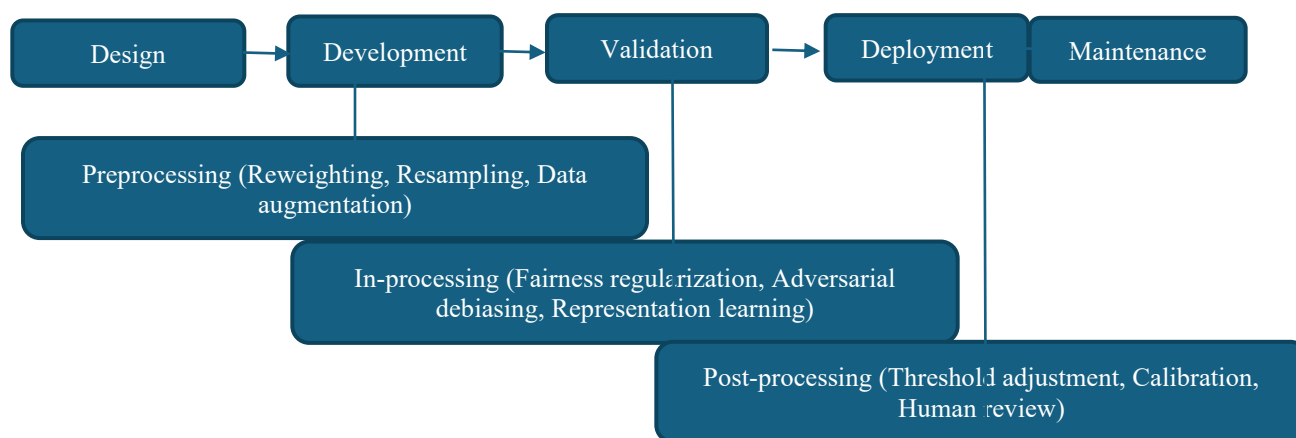


Figure 2. Three-Stage Mitigation Pipeline Mapped to the AI Assessment System Lifecycle

This lifecycle view matters because educational AI systems are rarely static. A writing scoring system may be retrained on new responses, a proctoring system may change its detection model, and a learning analytics tool may add new predictors, all without a vendor name change. A fairness evaluation conducted at purchase does not necessarily apply to a later version, so maintenance controls should include version tracking, change notices, retained evaluation reports, and a process for comparing pre-update and post-update performance. If a system expands from one course to an entire program, revalidation should be required because the deployment population and stakes have changed.

Vendor Opacity and Proxy Testing

Vendor opacity is the condition under which this framework most often has to operate in practice. Many deployed systems are proprietary, and an institution frequently cannot inspect training data, model architecture, or internal thresholds. Continuous monitoring language is not credible without specifying what an institution can do when full transparency is unavailable. Three techniques remain available even under vendor opacity. First, proxy (behavioral) testing, where an institution does not need model internals to submit matched or perturbed inputs, such as essays that hold argument structure constant while varying dialect features, or proctoring sessions that hold behavior constant while varying lighting and camera angle, and observe whether outputs change. This is the sensitivity test in Table 5 (Step 5), and it requires only input/output access. Second, output-level statistical monitoring, where an institution can track its own observed flag rates, score distributions, and appeal outcomes by subgroup over time even without model access, since these are properties of what the institution receives rather than of the vendor's code. Third, contractual audit rights, where procurement agreements can require vendors to supply the minimum evidence package (Table 7), disclose material model updates, and permit an independent third-party audit using the institution's own held-out cases. A vendor's refusal to agree to these terms is itself evidence

relevant to the risk-tier decision in Table 6. None of these substitutes for internal model access, but together they keep the monitoring obligations in this framework achievable rather than aspirational.

Boundary Conditions and Failure Modes

Boundary conditions describe the limits of a system's intended use. A writing model trained on short persuasive essays may not generalize to lab reports or multilingual portfolios, and a proctoring model calibrated for laptop webcams may not generalize to tablets or low-bandwidth connections. If a vendor cannot define these boundary conditions, the institution should assume the system's safe use is narrower than advertised, particularly where the system affects grades, academic integrity, or progression. Failure-mode analysis asks how the system fails and who absorbs the cost, classifying failures by visibility, reversibility, severity, and subgroup concentration, since that classification determines the appropriate response. A visible, reversible failure may warrant interface warnings and human review, while a hidden, systematic, subgroup-concentrated failure may warrant suspension until further evidence is available.

Minimum Evidence Package for Technical Approval

The framework recommends that any high-stakes AI assessment system produce the minimum evidence package in Table 7 before technical approval. These materials do not guarantee fairness, but they create the conditions for informed review, and their absence is itself a finding under the risk-tier policy in Table 6.

Evidence Element	What It Must Show
System description	What the tool does in operational terms, avoiding vague marketing language such as "AI-powered integrity protection."
Intended use / prohibited use statement	The educational settings and decision types the tool has evidence for, and those it does not - since many systems drift from advisory to decision-making use over time.
Dataset documentation	Source, time period, task type, language, demographic representation where available, and known gaps (Table 5, Step 2).
Validation results	Overall and subgroup performance, with sample sizes and confidence intervals (Table 5, Step 4).
Fairness metric justification	Why the selected criteria fit the assessment context and its stakes (Table 3).
Mitigation history	Which pre-, in-, or post-processing methods were used and their before/after subgroup effect (Table 4).
Explanation capability	Reason codes, confidence information, or evidence summaries appropriate to the actual reviewing audience - engineers, faculty, or students each need a different explanation.
Human review plan	When review occurs, what evidence is available, and whether the reviewer can override the system.
Monitoring schedule	When fairness evidence will be refreshed and what triggers early re-evaluation (Table 5, Step 8).

Table 7. Minimum Evidence Package for Technical Approval

In practice, many vendor systems will not initially meet this evidence standard. That does not make the standard unrealistic; it shows where current procurement practice is weak. Educational institutions are being

asked to trust systems that can influence student outcomes, and a minimum evidence package creates a basis for technical trust that is stronger than brand reputation, vendor assurance, or aggregate accuracy alone.

A Hypothetical Worked Example

The framework is illustrated below on a hypothetical technical review of an essay-scoring model, applied to the kind of public benchmark used elsewhere in this literature, such as the ASAP-AES and ELLIPSE corpora used by Fan and Yun (2026). This is a worked illustration of how the diagnostic procedure in Table 5 would be applied, and not an empirical evaluation. No model was trained or tested for this paper, and no performance figures provided are reported results. It is presented to show what the framework asks a reviewer to produce.

Construct clarification (Step 1) would require the reviewer to obtain the scoring rubric and confirm, before touching model outputs, whether the construct is holistic writing quality, argument quality specifically, or a composite of traits; because corpora of this kind draw essays from several state writing tests with different rubrics, this step would need to be repeated per prompt set rather than assumed uniform across the corpus. Dataset accounting (Step 2) would document the population and grade bands the corpus was drawn from and flag that a model trained on it should not be assumed valid for adult learners, or graduate admissions essays without separate validation. Label review (Step 3) would compare inter-rater reliability among the corpus's human raters by examinee subgroup, where subgroup labels are available, before treating human scores as ground truth. Subgroup performance evaluation (Step 4) would report scoring error by the subgroups available in the corpus metadata, such as English-learner status, with confidence intervals wide enough to reflect the corpus's actual per-subgroup sample sizes, and would refrain from a fairness claim in any subgroup too small to support one.

Sensitivity testing (Step 5) would use matched-essay pairs, of the kind Fan and Yun (2026) construct, that hold argument quality constant while varying L2-associated syntactic features, and would report whether the score changes as a result. Mitigation selection (Step 6) follows directly from what Step 5 finds: if the model's own latent representations encode linguistic-origin information, the in-processing representation-learning approach of Fan and Yun (2026) is a candidate; if the training corpus underrepresents a population the model will actually score, pre-processing resampling is a candidate instead. Deployment simulation (Step 7) would test negative cases the benchmark does not emphasize, such as essays with strong reasoning but nonstandard syntax. Monitoring (Step 8) would specify that any retraining on new prompt sets triggers Steps 3 and 4 again.

Carried through with an actual model and actual subgroup data, this sequence is the empirical validation this paper does not attempt; Table 5 is the protocol such a validation would follow, and is offered as a concrete starting point for the empirical work called for in the Limitations and Future Research Agenda section below.

Contribution to Information Systems Research

This paper contributes to information systems research in three ways. First, it applies a stated taxonomy-development method (Nickerson, Varshney, & Muntermann, 2013) to a domain-specific problem, producing a bias taxonomy whose categories are traceable to a documented development process rather than asserted informally. Second, it operationalizes algorithmic-fairness concepts into artifacts that a technical reviewer, developer, or procurement officer can act on directly. A context-to-metric mapping, a diagnostic and testing procedure, risk-based deployment tiers, and a minimum evidence package. These artifacts follow design-science conventions in that they are intended to be usable, testable, and revisable rather than purely descriptive. Third, it connects two literatures that IS-adjacent fairness work has often

treated separately. The machine-learning fairness literature and the educational-measurement tradition (AERA, APA, & NCME, 2014; Messick, 1989), showing that a system can satisfy conventional fairness metrics while still failing a measurement-validity standard specific to assessment.

The Bias Taxonomy and Mitigation Pipeline (BTMP) framework is intended to operate across the system lifecycle described in Figure 2, from design through maintenance, rather than as a one-time certification. Because AI assessment systems are frequently updated, expanded to new populations, or repurposed for higher-stakes uses than originally intended, a lifecycle-oriented framework is more useful to practitioners than a single-point audit. The framework does not resolve the impossibility results documented by Chouldechova (2017) and Kleinberg et al. (2017); instead, it makes the tradeoffs those results imply explicit and traceable, so that a technical reviewer can document which fairness criteria were prioritized, why, and under what conditions that prioritization should be revisited.

Limitations and Future Research Agenda

This paper is conceptual. The framework has not been applied to a live system, and the worked example in the previous section is explicitly hypothetical: it shows how the diagnostic procedure would be applied, not what it would find. No dataset, protected-attribute analysis, intersectional result, or mitigation-outcome comparison is reported here, and none should be inferred from the illustrative example. The taxonomy, metric mapping, and evidence package should be treated as a proposal to be tested empirically, not as a validated instrument. Readers seeking empirical evidence of bias in deployed systems should consult the primary studies cited throughout, particularly Wilson and Huang (2024), Yoder-Himes et al. (2022), Fan and Yun (2026), and Tan et al. (2026), each of which reports real findings on real systems or benchmarks.

A second limitation concerns the fairness metrics themselves. Most classification-based fairness metrics were developed for binary decisions and do not map cleanly onto the ordinal or continuous outcomes common in essay scoring, a limitation discussed above (Fairness Criteria and Tradeoffs) but not resolved by this paper. A third limitation concerns vendor opacity. The proxy-testing and contractual approaches proposed above (Lifecycle Controls, Boundary Conditions, and Vendor Opacity) are partial substitutes for genuine model access, and their reliability under adversarial or uncooperative vendor conditions is itself an open empirical question.

Several directions follow directly from these limitations. First, fairness metrics defined specifically for ordinal and continuous educational outcomes are needed, extending equal opportunity and equalized odds beyond their native binary-classification setting rather than forcing a threshold onto scores that were never meant to be binary. Second, an explicit crosswalk between this paper's three bias families and Suresh and Guttag's (2021) seven machine-learning harm sources would connect this domain-specific taxonomy to the general ML-fairness lifecycle literature and clarify where the two frameworks agree and diverge. Third, an empirical comparison of mitigation-technique combinations is needed that evaluates outcomes against educational-validity criteria, such as construct representation and correlation with independent measures of the intended skill, rather than against parity metrics alone, since a mitigation that improves parity while degrading validity has not solved the underlying problem. Fourth, black-box auditing methods for vendor-opaque systems deserve dedicated empirical study, including how reliably proxy testing on institution-held cases approximates what full model access would reveal. Finally, the illustrative worked example above should be carried out as an actual study, applying Table 5 to a public essay-scoring benchmark with real subgroup data and real mitigation comparisons, to test whether the diagnostic procedure surfaces the disparities that primary studies such as Wilson and Huang (2024) and Fan and Yun (2026) have already documented by other means.

Conclusion

AI-based educational assessment systems will continue to expand into essay scoring, proctoring, predictive analytics, and automated feedback. Ensuring that these systems support rather than undermine equitable learning environments requires more than good intentions; it requires a technical framework that treats fairness as a measurable, testable, and revisable engineering property. This paper has presented the Bias Taxonomy and Mitigation Pipeline (BTMP) framework, a bias taxonomy built with a stated development method, a context-to-harm-to-metric mapping differentiated by decision stakes, a three-stage mitigation pipeline, a consolidated diagnostic and testing procedure, risk-based deployment tiers, and a minimum evidence package for technical approval. The framework connects the machine-learning fairness literature to the older educational-measurement tradition, and it is offered to developers, vendors, and technical reviewers as a working tool rather than a finished certification standard. Its value will depend on empirical testing against real systems and real student data, which this paper identifies as the necessary next step rather than a task it has completed.

References

- Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., & Wallach, H. (2018). A reductions approach to fair classification. *Proceedings of the 35th International Conference on Machine Learning*, 80, 60-69.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.
- Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V.2. *The Journal of Technology, Learning and Assessment*, 4(3).
- Baker, R. S., & Hawn, A. (2022). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*, 32(4), 1052-1092. <https://doi.org/10.1007/s40593-021-00285-9>
- Balash, D. G., Fainchtein, R. A., Korkes, E., Grant, M., Sherr, M., & Aviv, A. J. (2023). Educators' perspectives of using (or not using) online exam proctoring. *Proceedings of the 32nd USENIX Security Symposium*.
- Breyer, F. J., Attali, Y., Williamson, D. M., Ridolfi-McCulla, L., Ramineni, C., Duchnowski, M., & Harris, A. (2014). A study of the use of the e-rater scoring engine for the analytical writing measure of the GRE revised General Test (Research Report No. RR-14-24). Educational Testing Service.
- Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research*, 81, 77-91.

- Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2), 153-163.
- Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214-226.
- Fan, K., & Yun, E. (2026). Mitigating bias in automated grading systems for ESL learners: A contrastive learning approach. *arXiv preprint arXiv:2601.16724*.
- Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, 29, 3315-3323.
- Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1), 1-33.
- Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. *Proceedings of the 8th Innovations in Theoretical Computer Science Conference*.
- Kusner, M. J., Loftus, J. R., Russell, C., & Silva, R. (2017). Counterfactual fairness. *Advances in Neural Information Processing Systems*, 30, 4066-4076.
- Loukina, A., Madnani, N., & Zechner, K. (2019). The many dimensions of algorithmic fairness in educational applications. *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 1-10. <https://doi.org/10.18653/v1/W19-4401>
- 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.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-104). American Council on Education and Macmillan.
- Nickerson, R. C., Varshney, U., & Muntermann, J. (2013). A method for taxonomy development and its application in information systems. *European Journal of Information Systems*, 22(3), 336-359. <https://doi.org/10.1057/ejis.2012.26>
- Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., & Weinberger, K. Q. (2017). On fairness and calibration. *Advances in Neural Information Processing Systems*, 30, 5680-5689.
- Suresh, H., & Gutttag, J. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. *Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '21)*, Article 17, 1-9. <https://doi.org/10.1145/3465416.3483305>

- Tan, M., Phalen, L., & Demszky, D. (2026). Marked pedagogies: Examining linguistic biases in personalized automated writing feedback. *Proceedings of the LAK26: 16th International Learning Analytics and Knowledge Conference*, 568-577. <https://doi.org/10.1145/3785022.3785113>
- Wilson, J., & Huang, Y. (2024). Validity of automated essay scores for elementary-age English language learners: Evidence of bias? *Assessing Writing*, 60, 100815. <https://doi.org/10.1016/j.asw.2024.100815>
- Yoder-Himes, D. R., Asif, A., Kinney, K., Brandt, T. J., Cecil, R. E., Himes, P. R., Cashon, C., Hopp, R. M. P., & Ross, E. (2022). Racial, skin tone, and sex disparities in automated proctoring software. *Frontiers in Education*, 7, 881449. <https://doi.org/10.3389/feduc.2022.881449>
- Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning fair representations. *Proceedings of the 30th International Conference on Machine Learning*, 28(3), 325-333.
- Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 335-340.