Interpreting Fairness Evidence in Automated Scoring Through Validity, Risk, and Detection

Anonymous Submission

Abstract

As NLP systems are increasingly used in high-stakes educational and professional assessment, subgroup disparities in model outputs are often treated as evidence of fairness concerns. Yet such disparities are difficult to interpret in isolation: they may reflect construct-relevant proficiency differences, construct-irrelevant influences introduced by the scoring system, or artifacts of evaluation design. This paper presents a critical review of fairness evidence in automated scoring and argues that fairness evaluation requires alignment among three elements: a validity claim specifying score meaning and use, a risk account identifying how construct-irrelevant variance may enter the scoring pipeline, and a detection contrast capable of evaluating the relevant subgroup difference. Synthesizing research from NLP, machine learning fairness, and educational measurement, we show that prior work often examines these elements in isolation, leading to fairness claims that are under-specified or exceed what the evidence can support. We analyze common detection baselines, identify three recurring patterns of misalignment, and articulate an alignment standard for making bounded, interpretable fairness claims in deployed NLP scoring systems.

1 Introduction

As NLP systems move from benchmark tasks to institutional decision-making, the field faces a methodological challenge: researchers have become increasingly skilled at detecting disparities in model behavior, but less clear about how those disparities should be interpreted as evidence of fairness concerns. This challenge is especially consequential in automated scoring, where NLP systems evaluate constructed response item types such as short-answer questions, essays, and speaking prompts to produce scores that may inform admissions, promotion, credentialing, and remediation decisions (Williamson et al., 2012; Madnani and Cahill, 2018; McCaffrey et al., 2022).

A substantial body of work has examined fairness in automated scoring. Studies have documented group differences in scoring accuracy, prediction error, false-positive and false-negative rates, calibration, and related performance measures by race, gender, language background, and other characteristics (Loukina et al., 2019; Johnson et al., 2022; Yang et al., 2024; Schaller et al., 2024). This literature has made disparities in automated scoring increasingly visible. However, the simple presence of a disparity does not indicate that an observed difference reflects a fairness issue or warrants changes to scoring models, training data, reporting practices, and/or deployment conditions.

This challenge reflects a broader critique within NLP fairness research. Blodgett et al. (2020) argue that observed group differences in model behavior cannot be interpreted independently of the specific populations, harms, and institutional contexts involved, and that terms such as "bias" and "fairness" are often used without sufficient conceptual specificity to determine whether a disparity reflects an expected difference, a meaningful problem, or a statistical pattern with limited practical significance. Related work has questioned whether commonly used fairness metrics adequately connect statistical disparities to substantive concerns. Dev et al. (2022) note that fairness is often operationalized through statistical criteria without a clear justification for why those criteria represent harm within a particular application. Systems may satisfy a given metric while still producing undesirable outcomes for some groups, or they may violate a metric in ways that are not substantively concerning for the system's intended use.

Additionally, researchers have emphasized that disparities may emerge from multiple stages of the machine learning pipeline and therefore cannot be interpreted independently of the processes that pro-

duce them. Suresh and Guttag (2021) demonstrate that unwanted bias can emerge during problem formulation, data collection, annotation practices, feature representation, model optimization, or deployment conditions, suggesting that the same disparity may reflect very different underlying mechanisms. Detecting a group difference therefore does not, on its own, identify the source of the disparity within the ML pipeline or enable its mitigation. Delobelle et al. (2024) argue that many NLP fairness evaluations provide limited guidance on the practical significance of observed group differences, the mechanisms that produce them, or the actions needed to address them.

These critiques suggest that fairness evaluation requires more than identifying statistical disparities between groups. Consistent with measurement-centered accounts of fairness (Jacobs and Wallach, 2021) and argument-based accounts of assessment validity (Kane, 2013), observed group differences should be interpreted within a broader validity framework that specifies what a system is intended to measure, how its outputs will be used, which populations may be affected, and what forms of construct-irrelevant variance may threaten intended score interpretations and uses. Construct-irrelevant variance refers to score variation not attributable to the construct that the assessment is intended to measure, and, when systematic, it can contribute to bias.

This paper argues that fairness evaluation for deployed NLP scoring systems is most interpretable and useful when three perspectives are considered. A validity framework defines intended score interpretations and uses, the relevant populations and comparison groups, and the consequences attached to score interpretations and uses. Risk accounts examine how scoring-system design choices can introduce or amplify construct-irrelevant variance relative to validity specifications. Empirical detection methods evaluate whether relevant group differences emerge in scoring outcomes.

Existing work has examined these perspectives separately. Research in educational measurement has emphasized validity arguments (Kane, 2013; Messick, 1998; Ferrara and Qunbar, 2022; Lane and Marion, 2025), while fairness research in NLP and machine learning has focused on statistical disparities, bias mechanisms, and harm detection (Hardt et al., 2016; Chouldechova, 2017; Blodgett et al., 2020; Suresh and Guttag, 2021). These perspectives are deeply interdependent, and detected disparities are difficult to interpret without a validity framework, difficult to diagnose without attention to technical mechanisms, and speculative without empirical detection. The goal of this review is to bridge these perspectives to examine fairness evidence in automated scoring.

Automated scoring provides a particularly useful domain for this analysis because it sits at the intersection of educational measurement, NLP, and responsible AI. Unlike many NLP applications, automated scoring has a well-developed validity tradition in which the inferential chain from observed performance data to score interpretations and uses is explicitly theorized (Kane, 2013; Williamson et al., 2012; Ferrara and Qunbar, 2022; Lane and Marion, 2025). Automated scoring systems increasingly rely on NLP methods whose behavior must be evaluated across diverse populations and operational contexts (Loukina et al., 2019; Johnson et al., 2022; Johnson and McCaffrey, 2023). This combination makes automated scoring an especially revealing case for examining how fairness evidence is constructed, interpreted, evaluated, and translated into claims about deployed NLP systems.

Recent work has developed validity arguments for AI-based automated scores and connected argument-based validation with Responsible AI principles for the design, evaluation, and governance of AI-powered assessments (Ferrara and Qunbar, 2022; Burstein and LaFlair, 2025). Our focus is narrower and more diagnostic: we ask what evidence is needed to interpret group disparities observed in automated scoring outcomes as evidence of fairness. This distinction matters because validity frameworks specify intended score interpretations and uses, but may not specify how automated scoring systems can introduce or amplify construct-irrelevant variance or the technical approaches needed to fix such issues. NLP fairness studies often detect group differences without fully specifying the validity argument or the plausible mechanisms by which construct-irrelevant variance may enter the scoring process. Johnson and McCaffrey (2023) link these traditions by emphasizing psychometric detection methods. We complement that work by showing how validity arguments, hypothesized sources of construct-irrelevant variance, and detection contrasts must align before a detected group difference can support a bounded fairness claim.

This review's contribution is therefore not a new taxonomy of bias sources or a new fairness metric.

We propose an alignment standard for fairness evidence in automated scoring: a fairness claim is well supported only when validity evidence, hypothesized risk mechanism(s), and empirical detection contrast all support the same interpretation. This framing clarifies why common fairness evaluations remain under-specified even when they report statistically significant group differences.

The remainder of the paper is organized as follows. Section 2 explains why fairness concerns in automated scoring depend on validity arguments. Section 3 examines where construct-irrelevant variance may enter or be amplified in scoring systems. Section 4 reviews empirical contrasts used to detect group differences and clarifies what those contrasts can and cannot establish. Section 5 synthesizes recurring patterns of misalignment and discusses how fairness in deployed NLP scoring systems should be interpreted, evaluated, and reported. Section 6 outlines the scope and limitations of the paper's argument-driven review and proposed alignment standard.

2 Fairness as a Validity Issue

In the context of educational measurement, fairness in automated scoring is fundamentally a validity issue (Lane and Marion, 2025). Validity can be thought of as the extent to which both theory and empirical evidence support the meanings assigned to assessment outcomes (e.g., scores, feedback) and the ways in which assessment outcomes are intended to be used (Kane, 2013; Lane and Marion, 2025). Consider a spoken-response assessment that learners use to practice their communication skills. In the assessment, learners respond verbally to virtual prompts and receive feedback based on an automatically transcribed and machine-scored response.

In this context, the score is intended to represent the targeted communication construct, and its intended use is formative: supporting reflection and skill development. Observed group differences in scores would not inherently offer evidence of unfairness. Some differences may appropriately reflect variation in the underlying construct; for example, learners with more relevant training may demonstrate stronger performance. Other differences may arise from construct-irrelevant factors. If transcripts for learners with certain accents contain more ASR errors and the scoring model assigns lower scores to those distorted transcripts, the resulting group difference would reflect transcription distortion rather than communication skills.

From a validity perspective, a fair test produces scores that mean the same thing for all test takers and that do not unexpectedly advantage or disadvantage one group over another (Lane and Marion, 2025). This aligns with argument-based approaches to validity, which conceptualize validation as the evaluation of a series of inferences linking observed test performance data to score interpretations and uses (Kane, 2013). Fairness concerns emerge when the evidentiary chain supporting score interpretations and uses does not hold equally across groups. The central question is therefore not simply whether group differences exist, but whether those differences are consistent with what scores are supposed to mean and how they are supposed to be used.

The consequences associated with score interpretations and uses are another important element of an argument-based approach to validity and validation (Kane, 2013; Lane and Marion, 2025). If scores are used only as formative feedback for practice and learning, small group differences may have limited practical consequences (Bennett, 2011). However, if the same scores are used for high-stakes decisions, even small construct-irrelevant differences may represent fairness concerns. Here then, fairness depends not only on whether the inferences linking assessment data to score interpretations and uses are sound, but also on the type and magnitude of the implications of those score meanings and applications.

Establishing fairness therefore begins before the automated scoring process (Johnson et al., 2022). Test developers must specify the target population, construct of interest, intended score interpretations and uses, and the groups for which score differences are and are not expected. These groups need not be limited to demographic categories commonly examined in automated-scoring fairness work, such as race, gender, or language background (Loukina et al., 2019; Johnson et al., 2022); they may also include educational pathway, training type, disability status, or other factors that shape how examinees interact with the assessment and the scoring system. Some group differences may reflect construct-relevant variation in training, experience, or learning opportunities, while others may indicate construct-irrelevant variance introduced by the scoring pipeline. The selection of comparison groups is therefore important, because fairness

claims grounded in group differences can only be evaluated relative to the score interpretations and uses specified in an assessment's validity argument (American Educational Research Association et al., 2014).

A validity framework shifts the evaluation of fairness in automated scoring systems away from isolated statistical parity metrics toward a broader inquiry into whether scores remain meaningful, defensible, and comparable across relevant groups. Fairness evidence is not limited to detecting group disparities, but instead contributes to an integrated validity argument regarding whether automated scoring systems support appropriate and equitable score interpretations and uses.

Within automated scoring systems, construct-irrelevant variance may emerge through multiple components of the scoring pipeline, including training data properties, linguistic representation choices, prompt design, model architecture, and/or deployment conditions (Suresh and Guttag, 2021). These mechanisms may cause scoring systems to rely on linguistic or contextual features that are correlated with demographic or educational characteristics but unrelated to what the assessment is trying to measure. The central fairness question therefore becomes whether automated scoring systems introduce or amplify construct-irrelevant variance in ways that distort intended score interpretations and uses or result in unexpected problematic consequences for particular groups.

3 Construct-Irrelevant Variance in Automated Scoring Systems

Fairness concerns in automated scoring are often evidenced by unexpected group score differences stemming from construct-irrelevant variance introduced or amplified by the system itself. Such variance may arise from demographic imbalance in training data, signals embedded in linguistic representations, prompt formulations that privilege particular language varieties, model-learning dynamics that amplify spurious correlations, or deployment conditions such as speech-recognition errors that differentially affect particular groups. We therefore ask: where in the scoring pipeline could construct-irrelevant variance enter, and through what mechanisms could it produce problematic group differences in outcomes?

Several influential works provide conceptual tools to identify potential sources of construct-irrelevant variance, including Suresh and Guttag (2021), Hovy and Prabhumoye (2021), and Shah et al. (2020). In the broader machine-learning literature, such sources are often discussed as forms of bias. In this section, we treat bias, fairness risk, and construct-irrelevant variance as related but not interchangeable terms. Our focus is specifically on construct-irrelevant variance: ways in which automated scoring systems may be influenced by factors unrelated to the intended construct and thereby create or amplify group differences.

Suresh and Guttag (2021) provide a broad lifecycle account of harm caused by biases in machine-learning systems. They reject the vague explanation that a system is biased because "the data are biased." Instead, they locate possible sources of harm across the ML lifecycle: historical bias, representation bias, measurement bias, aggregation bias, learning bias, evaluation bias, and deployment bias. Doing so enables systematic risk analysis and points toward appropriate mitigations.

Hovy and Prabhumoye (2021) focus on NLP, identifying five sources of bias: data, annotation, input representations, models, and research design. A model may learn not only the intended construct signal, but also secondary linguistic signals associated with particular groups (e.g. age, gender, dialect, native-speaker status): bias flows from demographic variation in language, through data and representation choices, into model behavior.

Shah et al. (2020) provide the most explicit link between sources of bias and observable output disparities. Their predictive-bias framework distinguishes four entry points along NLP pipelines – the source population (selection bias), the labels (label bias), the embedding space (semantic bias), and the model-fitting process (overamplification) – whose consequences are observed as either predicted outcome disparity or prediction error disparity.

These frameworks classify bias at different levels of abstraction – across the ML lifecycle, sources of bias in NLP systems, or predictive-bias mechanisms that lead to outcome and error disparities. Within automated scoring of textual content, we use these frameworks to organize selected sources of construct-irrelevant variance around the points at which they are most likely to enter or be amplified in scoring systems.

Data-related sources. Training-data imbalance, labeling bias, and domain mismatch. Construct-irrelevant variance enters when observed labels are distorted measurements of the target construct,

when the labeling process depends on group membership as well as construct, or when the training distribution differs from the deployment distribution; the learned scoring model inherits both sampling skew and group-conditional label distortion.

Model architecture and learning dynamics. Models, including transformer-based models such as DeBERTa and recent LLMs, may learn linguistic features unrelated to the construct (He et al., 2021). Optimization exploits any representation component that reduces training loss; when construct-irrelevant components correlate with group membership, that correlation can propagate into predictions.

Feature and prompt design. The input features, rubrics, prompts, exemplars, and instructions that shape what the scoring system attends to. Surface features such as response length or vocabulary may influence scores, and generative or LLM-as-judge scoring systems introduce additional operationalization choices involving prompt wording, rubric framing, exemplar selection, and scoring instructions (Johnson and Zhang, 2024; Huang et al., 2026; Wang et al., 2026). Construct-irrelevant variance is transmitted when this operationalization makes surface features, linguistic style, prompt-aligned behaviors, or exemplar-specific patterns more salient. The resulting scores reflect these choices rather than the construct itself.

Deployment context. Scoring conditions, including input modality and pre-processing. Construct-irrelevant variance emerges or is amplified when the scoring system is used in a particular operational setting: pre-processing such as ASR can introduce group-dependent transcription errors before the model sees the input (Koenecke et al., 2020), and post-processing such as thresholding or institutional use can transform small score differences into large differences in acceptance, placement, or feedback decisions, thereby threatening intended score interpretations even when the scoring model itself is well-calibrated on clean input.

Table 1 maps the four entry points to related bias frameworks and transmission mechanisms. Table 2 maps selected studies to the sources of construct-irrelevant variance they discuss. Features, data, and model architecture dominate this literature; deployment context is the least discussed, likely because it requires studying operational pipelines rather than offline corpora.

An entry point of construct-irrelevant variance is a candidate explanation, not an observed effect. Whether any of them produces a subgroup difference in a given system is an empirical question, which Section 4 takes up through methods used to detect such differences and their limits.

4 Empirical Detection of Group Differences

Detection is the empirical layer of fairness evaluation. In NLP and machine learning, it is often operationalized through subgroup comparisons, such as differences in accuracy, error rates, calibration, or related parity criteria (Hardt et al., 2016; Chouldechova, 2017). These comparisons have been criticized when treated as fairness claims without specifying harms, populations, or use contexts (Blodgett et al., 2020; Dev et al., 2022; Delobelle et al., 2024). The problem is sharper in automated scoring, where model outputs are assessment scores used to support claims about examinee performance (Kane, 2013; Williamson et al., 2012). In this setting, detection asks which empirical contrast makes an observed subgroup difference relevant to the scoring process rather than attributable only to ratings, proficiency distributions, task exposure, response characteristics, or downstream decision rules. The choice of contrast also shapes which construct-irrelevant variance sources from Section 3 the result can plausibly speak to and which remain underdetermined.

4.1 Detection baselines

Automated-scoring studies use several kinds of comparisons to clarify the system's role in observed group differences. These are analytic distinctions rather than mutually exclusive designs, and each baseline changes what can be inferred from an observed subgroup difference.

Human Scoring (HS). Automated scores are compared with human ratings to ask whether subgroup differences exceed, mirror, or diverge from those produced by raters. Formally, this baseline conditions on the human score Y as the reference and asks whether the automated score $\hat{Y}$ relates to Y in the same way across subgroups. For each group G and reference R_G (typically the pooled population or a comparison subgroup), define the human–machine discrepancy[1]

$$\delta_{\text{HM}}(G) := \mathbf{f}(\hat{Y}, Y \mid G) - \mathbf{f}(\hat{Y}, Y \mid R_G), \quad (1)$$

[1]We provide formal specifications of HS in full to fix notation; MP, ASC, and RS follow the same template with different conditioning sets, and their formal specifications appear in Appendix A.

where $\mathbf{f}$ is a chosen vector of functionals of the joint distribution $(\hat{Y}, Y) \mid G$–for example, the standardised mean difference (SMD), agreement coefficients, variance explained, or any combination. A subgroup A^*_{HM} is flagged when

$$A^*_{\text{HM}} = \{\, G : \nu(\delta_{\text{HM}}(G)) = 1 \,\}, \qquad (2)$$

with ν an indicator function encoding the auditor's flagging rule – a threshold, a significance test, or a substantive reading of the discrepancy vector. Loukina et al. (2019) operationalize f through overall score accuracy, overall score difference, and conditional score difference; Litman et al. (2021) adopt related definitions for text-evidence scoring, and Bridgeman et al. (2012) use f = SMD to identify country-level over- and under-scoring by e-rater.

This baseline is useful but not neutral. Human ratings may themselves contain rater bias (Amorim et al., 2018) – that is, $Y \not\perp G$ even after conditioning on the construct – and human-machine discrepancies can vary across demographic groups (Bridgeman et al., 2012). $\nu(\delta_{\text{HM}}(G)) = 1$ signals only that the model's relationship to human scores differs across groups, leaving open whether the deviation reflects model bias, rater bias, or both.

If raters are systematically biased against a subgroup, a model that does *not* reproduce that bias will appear less accurate and less group-comparable against the human baseline–its $\delta_{\text{HM}}(G)$ will be large *because* it fails to reproduce the rater bias. Agreement with human scores may therefore reward a model for reproducing rater bias, while divergence from human scores is not, on its own, sufficient evidence of model unfairness.

Matched Proficiency (MP). DIF-style analyses ask whether examinees with comparable proficiency receive different scores as a function of group membership (Holland and Wainer, 1993; Zwick, 2012). This baseline conditions on a proficiency estimate θ and flags when the automated score depends on group membership once θ is held fixed – $\hat{Y} \not\perp G \mid \theta$. Kwako et al. (2022, 2023) apply this logic to BERT-based automated speech assessment.

These comparisons shift detection from raw subgroup differences to conditional ones – whether examinees estimated to have similar proficiency θ receive different scores across groups – with interpretation contingent on how θ is operationalized.

Alternative Scoring Condition (ASC). Studies vary model family, training data, prompt setting, or representation to ask whether subgroup differences are sensitive to system-design choices. This baseline compares two automated scoring systems S_1 and S_2 applied to the same responses and flags when the per-group cross-system gap itself depends on G. A system change that shifts all groups' scores equally is a recalibration concern, so detection fires only when the system change affects groups differently. Schaller et al. (2024) and Yang et al. (2024) apply this logic to AES systems across model classes, training conditions, accuracy, fairness, and generalizability. Representation-level analyses ask a related question: whether demographic information remains recoverable from model representations (Kwako and Ormerod, 2024). This family of comparisons shows whether fairness findings are design-sensitive, but attributing a disparity to a specific choice requires one-dimension-at-a-time comparisons.

Response Similarity (RS). Following the individual-fairness principle that similar individuals should be treated similarly (Dwork et al., 2012), Doewes et al. (2022) apply this logic to AES by asking whether similar essays receive similar scores. This baseline conditions on a construct-relevant similarity relation sim_C on responses (e.g. through a similarity threshold s) and flags when the response-to-score mapping of construct-equivalent responses shows local inconsistency.This baseline rests on sim_C being approximately independent of G. If $\text{sim}_C \not\perp G$, the matched pairs are no longer construct-equivalent.

The four baselines differ in what they can attribute. HS and ASC are more informative for Features, Model, or Deployment sources; HS flags the discrepancy without identifying which produced it, while ASC can isolate a system change. MP and RS are more informative for whether comparable examinees or responses are treated differently, including under annotation-related variance, but their conditioning on θ or sim_C confounds detection when $\theta \not\perp G$ or $\text{sim}_C \not\perp G$

4.2 Detection Limits

A recurring problem in automated-scoring fairness research is not weak detection methods but overextended claims – the gap between detecting a difference and knowing what that difference supports. Each element of the detection configuration $(\mathbf{f}, R_G, B, \nu, s)$ embeds a substantive choice: $\mathbf{f}$ selects which aspects of the score distribution are

instrumented, R_G specifies the reference, B (Y, θ, S, or sim_C) fixes what is held constant, ν converts measured discrepancies into a flag, and s sets the granularity at which local consistency is tested. Two audits computing $\delta(G)$ on the same data can produce different A^* depending on these choices. Each limit below corresponds to how a choice in this configuration becomes consequential.

A first limit is that conclusions depend on the comparison and the target quantity – that is, on the choice of $\mathbf{f}$ and the conditioning set B. Loukina et al. (2019) show that fairness conclusions in automated scoring depend on the definition being used: under different comparisons different conclusions can be reached for the same system. Studies that report only SMDs may miss disparities visible in error rates or calibration, and vice versa. A single $\mathbf{f}$ cannot stand in for the broader question of whether $\hat{Y} \not\perp G \mid B$.

A second limit is that statistical significance does not determine practical importance. A small but statistically significant subgroup difference may be practically negligible if the score is used only for formative feedback, but consequential if a threshold triggers mandatory remediation or restricts access to clinical experiences. A non-significant result may be too uncertain to support a no-disparity conclusion. Kwako et al. (2023) illustrate good practice: they report a statistically significant L1-related DIF increase in one grade band, but note that the effect is practically small. DIF procedures in educational measurement have long treated flagging rules and minimum sample sizes as consequential design choices (Zwick, 2012), and the same concern applies to automated-scoring fairness evaluations.

A third limit is that thresholds and decision rules silently convert detection into interpretation: the flagging rule ν does normative work. Schaller et al. (2024), for example, use a fairness threshold to determine whether models are flagged as unfair. Such thresholds can be useful screening devices, but $\nu(\delta(G)) = 0$ only means "not flagged" by that particular ν – not "fair". ν is part of detection design, not a substitute for interpretation.

A fourth limit is the reference-class problem in matched-proficiency analyses. DIF-style detection requires a definition of comparable examinees (Holland and Wainer, 1993; Zwick, 2012). In automated scoring, the reference class may be defined using observed human scores, latent proficiency estimates, or model predictions; each choice supports a different interpretation. A particular risk arises when proficiency is estimated from the same response evidence used to produce the score under evaluation: it is possible that $\theta \not\perp G$ – and construct-irrelevant variance may enter both the score and the proficiency estimate, allowing the matched comparison to absorb part of the disparity it is meant to reveal.

Detection therefore makes a bounded contribution to a fairness claim. It shows whether subgroup differences appear under a particular empirical contrast, and which differences remain after the chosen baseline is applied. It does not establish what produced those differences or whether they matter.

5 Aligning Fairness Concerns with Validity, Risk, and Detection

Fairness studies in NLP often begin with an outcome: a model performs differently for different groups. In automated scoring the hard question is not whether there are group differences, but what any observed group differences mean. A group difference in scores, errors, calibration, or agreement may indicate a fairness problem, but it may also reflect biased human ratings, modeling artifacts, or a deployment condition such as ASR error. Such differences also could reflect expected performance ranges or result from an appropriate decision rule applied after scoring.

The fairness of an automated scoring model and its associated scores can only be understood in relation to intended score interpretations and uses. A validity framework specifies what scores are intended to convey, how they should be used, and what group differences would raise fairness concerns. Given intended score interpretations and uses, a risk account identifies plausible ways in which construct-irrelevant variance could enter the scoring pipeline, and a detection contrast evaluates whether relevant group differences appear in the data. Fairness evidence is strongest when the hypothesized risk and detection contrast are consistent with the validity argument and with each other. When inconsistencies occur, the result may still be useful for hypothesis generation, but should not be treated as a standalone fairness claim. For example, if a spoken-response assessment is intended to support the development of communication skills, and the differential quality of the ASR is the hypothesized risk, an aligned design would compare scoring outcomes under ASR and human-transcription

conditions rather than treat the score gap alone as evidence of the unfairness of the model or the assessment as a whole.

5.1 Three patterns of misalignment

The reviewed literature illustrates three patterns of misalignment that can arise when risk accounts and detection results are not grounded in a validity framework. The first is detection without intended score uses and interpretations: a group difference is reported, but the purpose of the assessment and how its scores are intended to be used are unknown, so the practical significance of the difference remains unclear. For example, Schaller et al. (2024) compare AES systems under full and skewed training data and apply a fairness threshold to flag models as unfair. The cross-system contrast is informative, but a threshold-based flag is difficult to interpret without specifying the score use in which the flagged difference would matter. A small continuous-scale difference might only become consequential after a cut score is applied in order to make a pass/fail decision about an examinee. Without knowing how assessment scores are meant to be used, the flag cannot tell us whether the difference is negligible in a formative setting or consequential in a high-stakes decision setting. This first pattern is overarching and as such, perhaps the most crucial. Consistency between risk accounts and detection results in the absence of a validity argument leaves little room for a clear evaluation of fairness.

The second pattern is *mechanism without an attribution-capable contrast*: a plausible source of construct-irrelevant variance is identified relative to what scores are intended to mean, but the empirical contrast cannot isolate it from alternatives. Kwako and Ormerod (2024) show that demographic information remains recoverable from the hidden representations of automated essay scoring models. This is evidence that demographic information is encoded at the representation level, but recoverability alone does not establish that the signal drives the scoring decision. A representation can encode an attribute without using it predictively. Closing that gap requires a contrast linking representation-level recoverability to score-level disparity.

The third pattern is *metric–claim mismatch*, where a study uses one metric to support a fairness claim that requires different evidence. For example, a study may find no group difference in average scores and treat that as evidence that the system is fair for a pass/fail decision. But average scores do not show whether one group is more likely to be pushed below the cut score, whether error rates differ, or whether comparable examinees receive comparable scores. Loukina et al. (2019) show that the same scoring system can support different fairness conclusions under different definitions. The misalignment occurs when the reported metric supports a narrow claim, but the paper presents it as a broader fairness verdict.

These patterns are not mutually exclusive, nor are they offered as frequency statements about the field; the point is not to fault individual studies, most of which contribute useful evidence within their stated scope. A group difference without a validity argument to back it is a signal whose meaning remains unclear; a hypothesized mechanism without an attribution-capable contrast is a plausible risk that has not been empirically isolated; and a single-functional result is a measurement of one aspect of the score distribution. None is a complete fairness claim on its own.

5.2 Reporting Fairness Claims

The alignment standard implies a corresponding reporting practice. A well-supported fairness claim should specify: (1) the construct and intended score use; (2) the comparison groups and why they are relevant to the validity argument; (3) the hypothesized source of construct-irrelevant variance, defined in relation to those validity specifications, and where it could enter the scoring pipeline; (4) the detection contrast used to evaluate the relevant subgroup difference; (5) the effect size and uncertainty of the observed difference; and (6) the limits of attribution supported by the contrast. A human–machine discrepancy speaks to the model's relationship to human ratings, not to the scoring system as a whole; a matched-proficiency contrast depends on the chosen proficiency estimate; and an Alternative Scoring Condition contrast localizes the disparity to a pipeline change, not necessarily to a single component. This framing does not require every study to explain every source of disparity. It requires the claim to match the evidence: subgroup differences are not final fairness verdicts, but evidence whose meaning depends on score use, comparison groups, hypothesized source of construct-irrelevant variance, and detection contrast. Bounded claims that make these dependencies explicit accumulate into a fairness argument; isolated contrasts do not.

6 Limitations

This paper offers a targeted, argument-driven review rather than a systematic review of fairness research in automated scoring, educational AI, or NLP more broadly. Our goal is not to provide exhaustive coverage of all relevant studies, but to use selected work across automated scoring, educational measurement, and NLP fairness to clarify how different forms of evidence support different kinds of fairness claims. As a result, the patterns we identify should be read as recurring claim-evidence problems observed in the literature, not as estimates of their relevance across the full field.

The alignment standard developed in this paper is also tied to a specific domain. It is motivated by automated scoring systems, where model outputs are interpreted as assessment scores and where validity arguments provide an established way to specify score meaning, intended use, and consequences. The broader logic of aligning validity claims, risk mechanisms, and detection contrasts may be useful for other NLP systems used in institutional decision-making, but the specific reliance on assessment validity is most direct in scoring contexts. In settings where the target construct is poorly specified or contested, alignment analysis may help make the disagreement more visible, but it cannot by itself resolve what the system ought to measure.

We do not empirically validate the alignment standard as an audit instrument, the misalignment patterns discussed in Section 5 are conceptual and illustrative rather than the result of systematic coding of a corpus of studies. A stronger empirical validation would require applying the standard to a larger set of automated-scoring fairness studies, examining whether independent reviewers reach convergent judgments about which claims are supported by which evidence, and testing whether aligned claims lead to more reproducible or actionable fairness conclusions across evaluation settings.

The risk sources and detection baselines discussed in this paper are intended as organizing distinctions for evaluating fairness claims, rather than as a closed taxonomy of all possible scoring-system failures. In practice, a single observed disparity may involve multiple parts of the scoring pipeline and empirical contrasts may draw on overlapping sources of evidence such as human ratings, proficiency estimates, model outputs, or decision thresholds. This does not undermine the alignment standard, rather, it reinforces the need to state which source of construct-irrelevant variance a study is targeting and what its chosen contrast can support. The same issue becomes especially important for LLM-as-judge, few-shot, and prompt-sensitive scoring systems, where prompt design, exemplar selection, model choice, and scoring instructions may each function as alternative scoring conditions.

Finally, the alignment standard clarifies what a detected disparity can support, but it does not determine what level of disparity is acceptable, who should set thresholds, or when intervention is required. Those decisions remain normative and institutional, even when the evidentiary basis for a fairness claim is made clearer.

7 Ethical Considerations

The alignment standard clarifies when a detected subgroup difference supports a bounded fairness claim, but it does not determine how such claims should be acted on. Three normative questions remain outside its scope: what level of residual disparity is acceptable, who has standing to specify the validity argument against which fairness is evaluated, and when intervention in a deployed scoring system is required. As noted in Section 6, these remain institutional decisions.

Two implications follow. First, validity arguments in automated scoring are typically specified by test developers and operators. The comparison groups, intended uses, and expected versus unexpected differences encoded in those arguments shape which disparities are treated as fairness-relevant. Detection contrasts aligned with such arguments may therefore inherit their omissions, and groups not represented in the validity argument may receive less scrutiny than groups it explicitly anticipates.

Second, the standard is diagnostic rather than corrective. It clarifies what a body of evidence supports, but does not by itself establish the full causal source of a disparity or prescribe a remedy. Bounded fairness claims should therefore be read as improving how disparity evidence is interpreted, not as resolving the construct-irrelevant variance that may have produced the observed disparity. A bounded claim is informative, but informativeness alone does not change the deployed scoring pipeline; identifying and reducing construct-irrelevant variance remains a separate task.

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. 2014. *Standards for Educational and Psychological Testing*. American Educational Research Association, Washington, DC.

Evelin Amorim, Márcia Cançado, and Adriano Veloso. 2018. Automated essay scoring in the presence of biased ratings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 229–237. Association for Computational Linguistics.

Perpetual Baffour, Tor Saxberg, and Scott Crossley. 2023. Analyzing bias in large language model solutions for assisted writing feedback tools: Lessons from the feedback prize competition series. In *Proceedings of the 18th workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 242–246.

Randy Elliot Bennett. 2011. Formative assessment: A critical review. *Assessment in education: principles, policy & practice*, 18(1):5–25.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476. Association for Computational Linguistics.

Brent Bridgeman, Catherine Trapani, and Yigal Attali. 2012. Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country. *Applied Measurement in Education*, 25(1):27–40.

Jill Burstein and Geoffrey T. LaFlair. 2025. Where assessment validation and responsible AI meet. *Language Teaching Research Quarterly*, 50:120–137.

Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153–163.

Pieter Delobelle, Giuseppe Attanasio, Debora Nozza, Su Lin Blodgett, and Zeerak Talat. 2024. Metrics for what, metrics for whom: Assessing actionability of bias evaluation metrics in NLP. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21669–21691. Association for Computational Linguistics.

Sunipa Dev, Emily Sheng, Jieyu Zhao, Aubrie Amstutz, Jiao Sun, Yu Hou, Mattie Sanseverino, Jiin Kim, Akihiro Nishi, Nanyun Peng, and Kai-Wei Chang. 2022. On measures of biases and harms in NLP. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 246–267. Association for Computational Linguistics.

Afrizal Doewes, Akrati Saxena, Yulong Pei, and Mykola Pechenizkiy. 2022. Individual fairness evaluation for automated essay scoring system. *International Educational Data Mining Society*.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226.

Steve Ferrara and Saed Qunbar. 2022. Validity arguments for AI-based automated scores: Essay scoring as an illustration. *Journal of Educational Measurement*, 59(3):288–313.

Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, volume 29.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Paul W. Holland and Howard Wainer, editors. 1993. *Differential Item Functioning*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Langdon Holmes, Wesley Morris, Scott Crossley, and Joon Suh Choi. 2026. Assessing fairness in finetuned scoring models with demographically restricted training data. *Assessing Writing*, 68:101032.

Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and linguistics compass*, 15(8):e12432.

Yue Huang, Corey Palermo, and Joshua Wilson. 2026. Accuracy and fairness of generative ai in automated essay scoring: Comparing gpt-4o, feature-based models, and human raters. *Assessing Writing*, 69:101047.

Abigail Z Jacobs and Hanna Wallach. 2021. Measurement and fairness. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 375–385.

Matthew Johnson and Mo Zhang. 2024. Examining the responsible use of zero-shot ai approaches to scoring essays. *Scientific Reports*, 14(1):30064.

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. In *Advancing Natural Language Processing in Educational Assessment*, pages 142–164. Routledge.

Michael T. Kane. 2013. Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1):1–73.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the national academy of sciences*, 117(14):7684–7689.

Alexander Kwako and Christopher Ormerod. 2024. Can language models guess your identity? analyzing demographic biases in AI essay scoring. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 78–86. Association for Computational Linguistics.

Alexander Kwako, Yixin Wan, Jieyu Zhao, Kai-Wei Chang, Li Cai, and Mark Hansen. 2022. Using item response theory to measure gender and racial bias of a bert-based automated english speech assessment system. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–7. Association for Computational Linguistics.

Alexander Kwako, Yixin Wan, Jieyu Zhao, Mark Hansen, Kai-Wei Chang, and Li Cai. 2023. Does bert exacerbate gender or l1 biases in automated english speaking assessment? In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 668–681. Association for Computational Linguistics.

Suzanne Lane and Scott F Marion. 2025. Validity and validation. *Educational measurement*, pages 191–275.

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 *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. Association for Computational Linguistics.

Nitin Madnani and Aoife Cahill. 2018. Automated scoring: Beyond natural language processing. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1099–1109. Association for Computational Linguistics.

Nitin Madnani, Anastassia Loukina, Alina Von Davier, Jill Burstein, and Aoife Cahill. 2017. Building better open-source tools to support fairness in automated scoring. In *Proceedings of the first ACL workshop on ethics in natural language processing*, pages 41–52.

Daniel F. McCaffrey, Jodi M. Casabianca, Kathryn L. Ricker-Pedley, René R. Lawless, and Cathy Wendler. 2022. Best practices for constructed-response scoring. *ETS Research Report Series*, 2022(1):1–58.

Samuel Messick. 1998. Test validity: A matter of consequence. *Social Indicators Research*, 45(1):35–44.

Ricardo Muñoz Sánchez, Simon Dobnik, Maria Irena Szawerna, Therese Lindström Tiedemann, and Elena Volodina. 2024. Did the names i used within my essay affect my score? diagnosing name biases in automated essay scoring. In *Proceedings of the workshop on computational approaches to language data pseudonymization (CALD-pseudo 2024)*, pages 81–91.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1668–1678.

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*, pages 210–221. Association for Computational Linguistics.

Deven Santosh Shah, H Andrew Schwartz, and Dirk Hovy. 2020. Predictive biases in natural language processing models: A conceptual framework and overview. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5248–5264.

Harini Suresh and John Guttag. 2021. A framework for understanding sources of harm throughout the machine learning life cycle. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. Association for Computing Machinery.

Manru Wang, Yihan Chen, Xiaoting Huang, and Yuxuan Lai. 2026. Opening the blackbox of llm-based automated essay scoring: Insights into feature weighting patterns and score validity. *Computers and Education: Artificial Intelligence*, page 100568.

David M. Williamson, Xiaoming Xi, and F. Jay Breyer. 2012. A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, 31(1):2–13.

Kaixun Yang, Mladen Raković, Yuyang Li, Quanlong Guan, Dragan Gašević, and Guangliang Chen. 2024. Unveiling the tapestry of automated essay scoring: A comprehensive investigation of accuracy, fairness, and generalizability. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22466–22474.

Rebecca J. Zwick. 2012. A review of ets differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement. *ETS Research Report Series*, 2012(1):i–30.

A Appendix: Formal Specifications of Detection Baselines

Each baseline in Section 4 can be written as a configuration $(\mathbf{f}, R_G, B, \nu, s)$ over a discrepancy functional δ and a flagging set A^*. The main text gives HS in full and states the conditioning structure, and briefly for MP, ASC, and RS. We provide the full specifications for the latter three here.

A.1 Matched Proficiency (MP)

For each group G and reference R_G, define the matched-proficiency discrepancy

$$\delta_{\text{MP}}(G \mid \theta) := \mathbf{f}(\hat{Y} \mid G, \theta) - \mathbf{f}(\hat{Y} \mid R_G, \theta), \quad (3)$$

where $\mathbf{f}$ is a chosen vector of functionals of the conditional distribution $\hat{Y} \mid (G, \theta)$ – a conditional mean, a regression coefficient on G after partialling out θ, or a more general distributional summary. A subgroup is flagged when

$$A^*_{\text{MP}} = \{\, G : \nu(\delta_{\text{MP}}(G \mid \theta)) = 1 \text{ for typical } \theta \,\}, \quad (4)$$

with ν encoding the auditor's flagging rule.

A.2 Alternative Scoring Condition (ASC)

For each group G and reference R_G, define the per-group cross-system discrepancy

$$\delta_{\text{ASC}}(G; S_1, S_2) := \mathbf{f}(\hat{Y}_{S_1} \mid G) - \mathbf{f}(\hat{Y}_{S_2} \mid G), \quad (5)$$

and the across-group double difference

$$\Delta_{\text{ASC}}(G; R_G, S_1, S_2) := \delta_{\text{ASC}}(G; S_1, S_2) - \delta_{\text{ASC}}(R_G; S_1, S_2). \quad (6)$$

A subgroup is flagged when

$$A^*_{\text{ASC}} = \{\, G : \nu(\Delta_{\text{ASC}}(G; R_G, S_1, S_2)) = 1 \,\}. \quad (7)$$

The double-difference structure ensures detection fires only on the fairness-relevant signal: a system change that shifts all groups' scores equally cancels in Δ_{ASC}.

A.3 Response Similarity (RS)

In RS, $\mathbf{f}$ is a Lipschitz-violation functional of $\hat{Y}$ over construct-equivalent pairs. Let $\mu_{G,s}$ denote the distribution of $(\hat{Y}(x_i), \hat{Y}(x_j), \text{sim}_C(x_i, x_j))$ induced by sampling $(x_i, x_j) \in \mathcal{P}_s(G)$, and define

$$\mathbf{f}(\hat{Y} \mid G, s) := \mathbf{E}_{\mu_{G,s}}\Big[\, \phi\big(|\hat{Y}(x_i) - \hat{Y}(x_j)|, \text{sim}_C(x_i, x_j)\big)\Big]. \quad (8)$$

where ϕ penalises score discrepancy in excess of what construct similarity permits. Doewes et al. (2022) take $\phi(d, c) = \mathbf{1}[d > 1 - c]$ and report $\mathbf{E}[\phi]$ pooled over the full pair set $\mathcal{P}$, without group conditioning or thresholding. Other choices recover related summaries: $\phi(d, c) = \max(0, d - (1 - c))$ gives mean violation magnitude, and $\phi(d, c) = d/(1 - c)$ a ratio summary. The construction extends naturally to subgroup flagging by restricting $\mu_{G,s}$ to within-group pairs, which permits group-level comparison, and imposing a similarity threshold s, which restricts attention to pairs where the Lipschitz inequality is non-trivial and stays closer to the principle "similar responses get similar scores":

$$\delta_{\text{RS}}(G) := \mathbf{f}(\hat{Y} \mid G, s) - \mathbf{f}(\hat{Y} \mid R_G, s), \quad (9)$$

$$A^*_{\text{RS}} = \big\{\, G : \nu(\delta_{\text{RS}}(G)) = 1 \,\big\}. \quad (10)$$

B Appendix: Potential sources of construct-irrelevance variance

Table 1: Proposed sources where construct-irrelevant variance can enter the pipeline, their transmission logic, and mapping to other frameworks

Construct-irrelevant variance category	**Automated scoring examples**	**Transmission logic**	**Mapping to other frameworks**
Data-related sources	Training-data imbalance; labeling bias; domain mismatch	Construct-irrelevant variance enters through the composition of the training or evaluation data, the human judgments used as labels, or the fit between the training domain and the target scoring context. In the source setting, the system learns from $P_s(Y_{\text{obs}} \mid X)$, where the observed label Y_{obs} is a potentially distorted measurement of the target construct Y^*, generated from a source-domain labeling process that may depend both on the underlying construct and group membership G: $Y_{\text{obs}} \sim P_s(Y_{\text{obs}} \mid Y^*, G)$. When $P_s \neq P_t$, the learned predictor $Q(\hat{Y} \mid X)$ may inherit both sampling skew and group-conditional label distortion, producing group-conditional differences in $Q(\hat{Y} \mid G)$.	Representation bias, measurement bias (Suresh and Guttag, 2021); data bias, annotation bias, (Hovy and Prabhumoye, 2021); selection bias and label bias (Shah et al., 2020).
Model architecture and learning dynamics	Transformer models such as DeBERTa; representation learning; spurious correlations; sensitivity to construct-irrelevant linguistic features	Construct-irrelevant variance is transmitted through what the model learns from the available data. The model may encode group-associated linguistic patterns, amplify correlations present in training data, or treat language features unrelated to the intended construct as predictive of score. Let the learned representation decompose as $Z = (Z_c, Z_g)$, where Z_c captures construct-relevant content and Z_g captures construct-irrelevant content correlated with group membership ($Z_g \not\perp G$). The optimization objective makes no distinction between Z_c and Z_g; any component of Z that reduces training loss is exploited, so the predictor Q depends on Z_g as well as Z_c. Group dependence in Z_g ($Z_a \not\perp A$) therefore induces group dependence in predictions, $\hat{Y} \not\perp G$, and where this dependence exceeds what is explained by true performance, $\hat{Y} \not\perp G \mid Y$.	Aggregation bias and learning bias (Suresh and Guttag, 2021); input-representation bias and model bias (Hovy and Prabhumoye, 2021); semantic bias and model overamplification (Shah et al., 2020).
Feature and prompt design	Surface features such as length and vocabulary; rubric wording; prompt dependence in generative scoring systems	Construct-irrelevant variance is transmitted through how the scoring construct is operationalized. Features, rubrics, prompts, or examples may make some aspects of a response more salient than others, causing the system to reward surface features, linguistic style, or prompt-aligned behaviors rather than the intended construct alone. Let ϕ denote the operationalization - features, rubric, or prompt - through which the construct C is rendered scorable, and decompose $\phi(X) = (\phi_c, \phi_s)$, where ϕ_c reflects construct-relevant content and ϕ_s reflects surface features made salient by ϕ. The scoring process makes no distinction between ϕ_c and ϕ_s; any component of $\phi(X)$ that contributes to the predicted score is exploited. Group dependence in ϕ_s ($\phi_s \not\perp G$) therefore induces group dependence in predictions, $\hat{Y} \not\perp G$, and where this dependence exceeds what is explained by true performance, $\hat{Y} \not\perp G \mid Y$.	Measurement bias (Suresh and Guttag, 2021); research-design bias, input-representation bias, and model bias (Hovy and Prabhumoye, 2021); semantic bias(Shah et al., 2020).
Deployment context	Speech versus text input; ASR preprocessing; subgroup differences in transcription errors; downstream use of model outputs	Construct-irrelevant variance emerges or is amplified when the scoring system is used in a particular operational setting. Input modality, preprocessing, transcription quality, post-processing, or downstream interpretation may affect groups differently, so the observed disparity may reflect the deployed scoring pipeline rather than the scoring model alone. The deployed system is $D = h(Q(\pi(X_{\text{raw}})))$: a pre-processing step π (e.g. transcription), the scoring model Q, and a post-processing step h (thresholding, banding, institutional use). Group-dependent error in π or non-linearity in h propagates into $D \mid G$ even when Q is well-calibrated on clean input.	Evaluation bias and deployment bias (Suresh and Guttag, 2021); research-design bias (Hovy and Prabhumoye, 2021); outcome disparity, and error disparity (consequence) (Shah et al., 2020).

Table 2: Studies discussed main sources of construct-irrelevant variance.

Study	Main source of construct-irrelevant variance	Key contributions/insights regarding sources of construct-irrelevant variance
Bridgeman et al. (2012)	Features	Features that capture organization and development carry the highest weight in e-rater, and "length is highly correlated with the combination of the organization and development features, which are the features with the greatest weight".
Madnani et al. (2017)	Data, Features, Model	A tool to detect features that could be a source of bias. training data, the features, and the learner function "evaluate the performance as well as the fairness of the entire scoring pipeline (the training data, the features, and the learner function)"
Sap et al. (2019)	Data	"...uncover unexpected correlations between surface markers of African American English (AAE) and ratings of toxicity in several widely-used hate speech datasets. Then, we show that models trained on these corpora acquire and propagate these biases..."
Koenecke et al. (2020)	Data, Deployment	"...we find evidence of a gap in the acoustic models but not in the language models."
Litman et al. (2021)	Data, Features	Word-count features were identified as a specific source. "Word count is not a fair feature". Removing word-count features reduced gender bias but also reduced scoring accuracy.
Kwako et al. (2022)	Model	"Transformer-based models can propagate and, in some cases, exacerbate gender and racial prejudice."
Baffour et al. (2023)	Data	Subjective or label-sensitive tasks may transmit annotation or construct ambiguity into subgroup performance differences.
Kwako et al. (2023)	Model	"Studies have revealed that pretrained LLMs can propagate and, in some cases, amplify negative stereotypes of marginalized groups."
Johnson and Zhang (2024)	Model	"We cannot rule out the possibility that GPT-4o infers, captures, and uses (construct-irrelevant) race/ethnicity information from the essays for the scoring task." Zero-shot GPT-4o predicted race/ethnicity from essays more accurately than it predicted human scores, and produced biased scores against Asian/Pacific Islander students after adjusting for performance.
Kwako and Ormerod (2024)	Model	"We found evidence that these group differences were embedded in the hidden layer of the model."
Sánchez et al. (2024)	Features, Model	"A feature-based system that does not explicitly use proper names should not exhibit name-based biases, while a model based on distributional semantics might pick up unwanted biases during its pre-training along all of the useful semantic information."
Schaller et al. (2024)	Data, Model	"the importance of including samples of the full range of users in the training data ... not only for demographic background variables but also for psychological aspects such as cognitive abilities"
Yang et al. (2024)	Model	Prompt-specific models improve accuracy but may learn prompt- or group-specific patterns that reduce fairness, especially for economic-status groups. Cross-prompt models test whether model behavior generalizes across task distributions.
Huang et al. (2026)	Data, Features, Model	Feature-based AES showed bias for gender and ELL groups, attributed to linguistic and stylistic tendencies varying across demographic groups in the training data. Prompt-based GPT-4 showed overall and residual subgroup bias, attributed to few-shot examples not linguistic and/or developmental diverse. Fine-tuned GPT-4 reached fairness 'nearly identical to that of human raters' for gender and ELL subgroups.
Holmes et al. (2026)	Data	Finetuning on demographically restricted training sets produced small systematic biases; balanced training data largely eliminated them. "[Our] LLMs function primarily as sophisticated pattern-matching systems that learn statistical regularities in human scoring behavior... any historical biases or measurement biases present in human ratings will likely be learned and reproduced by LLM-AES systems."
Wang et al. (2026)	Features, Model	"- LLM-based scoring, ..., shows greater sensitivity to different proficiency levels, with more pronounced weighting of stylistic and structural features as proficiency changes."