

Schema-Responsive Rubric Design for Epistemic Justice: A Fairness-Aware Assessment Framework for Multilingual Learners in Computer-Delivered English Proficiency Testing

Jaya Sinsinwar

Department of English, Jagannath University, Jaipur, Rajasthan, India

Published October 2026

Abstract

Computer-delivered English proficiency tests frequently penalize multilingual learners whose responses deviate from monolingual academic norms, thereby perpetuating epistemic injustice through rubrics that conflate communicative resourcefulness with linguistic deficit. We propose a schema-responsive rubric interface that mediates between learners' epistemic resources and assessment outcomes across four interconnected sub-modules. The Epistemic Justice Criteria Module replaces deficit-oriented descriptors with translanguaging recognition and oral narrative structure validation, wherein mixed-language segments are evaluated by pragmatic function rather than grammatical purity, and non-linear storytelling traditions are accommodated through adjusted coherence scoring. The Dynamic Rubric Rendering Engine then visually links response features to epistemic justice descriptors via context-aware highlighting, while an interactive calibration slider allows raters to modulate emphasis on communicative effectiveness against grammatical accuracy. Furthermore, the Rater Cognitive Support Layer embeds mandatory bias mitigation prompts and a standardized exemplar library to reduce inter-rater variability. The Audit and Feedback Loop subsequently ensures accountability through tamper-proof decision logging and generates feedback that explicitly validates recognized epistemic practices. The novelty of this framework lies in its integration of sociolinguistic theory with human-computer interaction design, transforming the rubric from a static judgment instrument into an adaptive mediator of fairness. Significance is twofold: methodologically, it offers a replicable architecture for embedding epistemic justice into assessment interfaces; practically, it fosters learner agency by acknowledging community-referenced reasoning and local idioms as legitimate evidence of proficiency.

Keywords: epistemic justice; schema-responsive rubric design; fairness-aware assessment; multilingual learners; computer-delivered English proficiency testing

1 Introduction

Computer-delivered English language proficiency assessments have become a dominant mode of measuring communicative competence in academic, professional, and migration contexts [1]. Their scalability and standardization offer clear administrative advantages, yet these same properties raise persistent questions about whether the scoring rubrics embedded in such systems adequately recognize the full range of multilingual learners' communicative repertoires [2]. Traditional analytic and holistic scoring approaches have long been critiqued for their inability to capture complex linguistic practices that fall outside monolingual academic norms [3]. Construct-centered assessment frameworks have attempted to align scoring with explicit theoretical constructs, but they frequently remain anchored to native-speaker benchmarks that treat deviation as deficiency [4].

A fundamental challenge confronting these systems is rater bias and cognitive load. Research on rater cognitive processes has demonstrated that human raters do not mechanically apply scoring rules; instead, they engage in complex mental modeling that is susceptible to implicit biases and contextual influences [5]. Rater calibration training and standardization procedures have been developed to reduce variability through exemplar-based learning [6]. Nevertheless, such methods often fail to address structural inequities because they calibrate raters toward a monolingual norm rather than interrogating the norm itself. The concept of epistemic injustice, originally formulated in philosophy, has recently been adapted to educational contexts to describe how marginalized groups are wronged in their capacity as knowers [7]. In language testing, this manifests as the misrecognition of non-standard varieties, translanguaging practices, and culturally specific discourse structures as deficits rather than resources [8].

The recognition of translanguaging as a legitimate communicative practice challenges the monolingual ideologies embedded in many scoring rubrics [9]. Work in multilingualism and literacy studies argues for the validation of hybrid language practices that draw on multiple linguistic resources simultaneously [10]. Furthermore, narrative inquiry and discourse analysis provide frameworks for understanding how oral narrative structures differ from written academic conventions, suggesting that standard rubrics may systematically penalize culturally specific storytelling modes [11]. These insights collectively point to a need for assessment interfaces that can accommodate diverse epistemic practices without sacrificing psychometric rigor.

Technologically, the proposed framework draws upon human-computer interaction research on fairness-aware design and algorithmic accountability [12]. Dynamic rendering engines relate to adaptive user interfaces and context-aware computing, which adjust information presentation based on user state or task complexity [13]. The integration of reflection prompts borrows from metacognitive scaffolding and reflective practice models used in teacher education and clinical decision support systems [14]. Additionally, the audit and feedback loop aligns with explainable AI principles, specifically feature attribution and decision trace logging, which aim to make semi-automated decisions transparent and contestable [15]. Foundational work in psychometrics also informs the need for validity arguments, particularly those concerning consequential validity, which examines the social impact of test scores [16]. The move toward equity-centered design in assessment emphasizes co-designing tools with affected communities to ensure they reflect diverse epistemic practices [17]. Cognitive load theory underpins the design of the rater support layer, aiming to optimize interface elements to prevent overload while enhancing decision quality [18].

We propose a Schema-Responsive Rubric Interface that embeds epistemic justice principles directly into the human rater interface of computer-delivered English language assessments. The core novelty lies in shifting the assessment framework from purely psychometric fairness to epistemic inclusion, where diverse knowledge systems and communicative repertoires are recognized as legitimate resources rather than deficits. Unlike prior work that treats rater training as a separate, pre-assessment activity, our approach integrates bias mitigation, dynamic rubric rendering, and decision trace logging into a unified interface that operates during scoring. The system comprises four interconnected modules: an Epistemic Justice Criteria Module that defines scoring dimensions through linguistic and discourse-functional rules; a Dynamic Rubric Rendering Engine that provides context-aware visual guidance and interactive calibration tools; a Rater Cognitive Support Layer that employs mandatory reflection prompts and exemplar libraries; and an Audit and Feedback Loop that ensures accountability by logging decision traces and generating learner-centric feedback. This

approach operationalizes abstract justice concepts into concrete, actionable user interface elements and scoring algorithms, aiming to reduce misrecognition of multilingual learners' performance.

The contribution of this work is threefold. First, we offer a theoretical bridge between epistemic justice philosophy and assessment interface design, demonstrating how abstract concepts such as testimonial and hermeneutical injustice can be translated into specific scoring criteria and interaction patterns. Second, we present a replicable architecture that integrates sociolinguistic theory with human-computer interaction, transforming the rubric from a static judgment instrument into an adaptive mediator of fairness. Third, we provide empirical evidence from a mixed-methods evaluation showing that the proposed interface reduces misrecognition of translanguaging and oral narrative structures while maintaining acceptable inter-rater reliability. These contributions distinguish our work from existing fairness-aware assessment approaches that focus primarily on statistical bias correction rather than epistemic inclusion.

The remainder of this paper reviews related work in language assessment, fairness-aware design, and epistemic justice; presents the theoretical foundations that inform our framework; describes the design and architecture of the Schema-Responsive Rubric Interface; reports the experimental evaluation; and discusses implications and future directions and concludes the paper.

2 Related work

Research on fairness in language assessment has expanded considerably over the past decade, yet the majority of existing approaches address bias through statistical correction or rater training rather than through the redesign of the assessment interface itself. This section reviews four strands of scholarship that inform our work: algorithmic fairness and bias mitigation in educational AI, epistemic injustice in assessment contexts, translanguaging and multilingual assessment, and human-computer interaction for fair decision-making. We then position our proposed framework against these existing efforts.

2.1 Algorithmic fairness and bias mitigation in educational AI

A substantial body of work has examined how bias propagates through AI-mediated educational systems and how fairness-aware algorithms can mitigate its effects. Recent scholarship has traced the migration of bias across epistemic layers, from human cognition to predictive models, arguing that fairness must be treated as an explicit optimization objective rather than an afterthought [19]. In the context of generative AI, feminist epistemological perspectives have been employed to interrogate how technological reproduction reinforces genealogies of inequality, particularly in recruitment and human-resource development [20]. These contributions highlight that fairness-aware algorithms, while useful, cannot alone resolve the deeper epistemic imbalances embedded in training data and model architectures.

Within educational assessment specifically, critical syntheses of AI-powered student performance modeling have raised concerns about algorithmic personalization and its epistemological consequences [21]. The central insight from this literature is that predictive accuracy is insufficient as a criterion for educational fairness; the conceptions of justice held by system designers fundamentally shape what counts as a fair outcome. Our framework responds to this concern by embedding justice principles directly into the scoring interface, rather than relegating them to post-hoc statistical adjustment.

2.2 Epistemic injustice in assessment and institutional communication

The concept of epistemic injustice has been productively extended to institutional communication and assessment. Work on algorithmic fairness in AI-mediated institutional communication has developed methodologies that integrate discourse analysis with computational frameworks to examine how institutional authority and epistemic control are exercised through language [22]. This line of inquiry demonstrates that fairness-aware machine learning must account for the discursive dimensions of epistemic power, not merely the distributional properties of data.

Related work has examined how linguistic markers such as epistemic downtoners and confirmatory tags signal varying degrees of certainty, and how these markers interact with fairness-aware governance of AI systems [23]. The theoretical foundations linking sociolinguistics to computational fairness have been articulated through research on discourse analysis and fairness-aware machine learning, which shows that categorical statements and confirmatory tags may signal high epistemic certainty in ways that automated systems fail to recognize [23]. Furthermore, work on reframing AI bias for equitable and culturally responsive education has argued that addressing epistemic injustice within AI systems requires participatory development and inclusive pedagogical frameworks [24]. These studies collectively establish that epistemic injustice in assessment is not merely a philosophical concern but a concrete design problem with computational implications.

2.3 Translanguaging and multilingual assessment

The validation of translanguaging practices in assessment has been a focus of sociolinguistic and educational research. Studies of linguistic justice in digital spheres have shown that claims of algorithmic neutrality are both epistemically and politically violent when they ignore the multilingual realities of speakers [25]. This work emphasizes that tools drawing their training data predominantly from English-language sources create epistemic imbalances that disadvantage non-dominant language users. The development of fairness-aware speculative decoding and related techniques represents one technical response, but such methods operate at the level of model output rather than assessment design.

Research on expanding methodological boundaries in language assessment has similarly noted that epistemic imbalance manifests concretely in tools whose training data are drawn from English-language sources [26]. Our framework addresses this imbalance by recognizing translanguaging as a legitimate epistemic resource within the rubric itself, rather than treating it as a deviation to be corrected.

2.4 Human-computer interaction for fair decision-making

The design of interfaces that support fair human decision-making has been informed by research on explainable AI, cognitive load, and reflective practice. Feature attribution methods and decision trace logging have been proposed to make semi-automated decisions transparent and contestable [15]. Cognitive load theory provides guidance on optimizing interface elements to prevent overload while enhancing decision quality [18]. Metacognitive scaffolding and reflective practice models, originally developed in teacher education and clinical decision support, offer precedents for embedding reflection prompts into professional workflows [14]. Our Rater Cognitive Support Layer draws on these traditions but adapts them specifically to the epistemic justice concerns that arise in multilingual assessment.

The proposed Schema-Responsive Rubric Interface differs from existing fairness-aware assessment approaches in several key respects. First, whereas prior work focuses on statistical bias correction or pre-assessment rater training, our framework integrates bias mitigation, dynamic rubric rendering, and decision trace logging into a unified interface that operates during scoring. Second, existing approaches to translanguaging in assessment tend to treat it as a separate construct to be measured, whereas we embed translanguaging recognition directly into the scoring criteria through pragmatic function analysis. Third, our framework operationalizes epistemic justice concepts into concrete user interface elements and scoring algorithms, providing a replicable architecture that bridges sociolinguistic theory and human-computer interaction design. These distinctions position the proposed work as a novel contribution that moves beyond both purely technical fairness interventions and purely theoretical critiques of assessment bias.

3 Theoretical foundations: epistemic justice and multilingual assessment

The design of a schema-responsive rubric interface rests on a set of theoretical commitments that must be made explicit before any architectural or algorithmic detail can be justified. This section develops those commitments in formal terms, beginning with a precise characterization of epistemic injustice as it arises in scoring, proceeding to a functional account of translanguaging and narrative structure as epistemic resources, and concluding with a formalization of the rubric as a mapping whose properties can be analyzed. Throughout, the objective is to establish the conditions under which an assessment interface can be said to be epistemically just, and to derive the constraints that any such interface must satisfy.

3.1 Formalizing epistemic injustice in scoring

Epistemic injustice, in the sense relevant to assessment, occurs when a learner is wronged in their capacity as a knower, either because their testimony is discounted due to identity prejudice (testimonial injustice) or because the interpretive resources available to make sense of their experience are absent (hermeneutical injustice) [7]. To render this concept tractable for interface design, we model the scoring process as a transformation from a response to a score, and we identify the points at which injustice can enter.

Let R denote the space of possible learner responses, and let $S \subset R$ denote the score space. A rubric induces a scoring function $f : R \rightarrow S$. In conventional assessment, f is composed of a feature extraction map $\phi : R \rightarrow \mathbb{R}^d$ and a weighting vector $w \in \mathbb{R}^d$, so that the score is given by the inner product

$$f(r) = \langle w, \phi(r) \rangle + b \quad (1)$$

where $b \in \mathbb{R}$ is an intercept term and d is the number of scored dimensions. Each coordinate $\phi_j(r)$ represents the measured value of dimension j for response r , and each w_j represents the weight assigned to that dimension.

Within this formulation, testimonial injustice corresponds to a systematic downward distortion of $\phi_j(r)$ for responses produced by learners from marginalized groups, even when the underlying communicative competence is equivalent. Formally, let $g \in G$ denote a learner's group membership, where G is a set of social categories. Testimonial injustice is present when there exists a dimension j and two responses r, r' with equivalent communicative content such that

$$E[\phi_j(r) \mid g = a] < E[\phi_j(r') \mid g = b] \quad (2)$$

for distinct groups $a, b \in G$, where the inequality reflects prejudice rather than genuine differences in competence. Hermeneutical injustice, by contrast, corresponds to the absence of a dimension in the feature map: if a communicative practice is not represented in ϕ , then no weighting scheme can recover it. Formally, hermeneutical injustice occurs when there exists a communicative resource c such that c is systematically deployed by group a but $\partial \phi_j / \partial c = 0$ for all j , meaning that the resource is invisible to the scoring function.

These two formalizations yield distinct design obligations. Testimonial injustice requires that the feature extraction map be calibrated so that equivalent competence yields equivalent feature values across groups, which motivates the bias mitigation prompts and exemplar calibration in the Rater Cognitive Support Layer. Hermeneutical injustice requires that the feature map be expanded to include dimensions that capture currently unrecognized resources, which motivates the Epistemic Justice Criteria Module. The schema-responsive interface is therefore not merely a presentation layer but a mechanism for modifying ϕ and w in response to detected epistemic resources.

3.2 Translanguaging as a functional resource

The recognition of translanguaging as a legitimate communicative practice requires a functional rather than a formal criterion for evaluating mixed-language segments [9]. In formal terms, consider a response r segmented into a sequence of units $u_1, u_2, \dots, u_n$, where each unit may be drawn from one or more linguistic codes. Let $l(u_i) \subseteq L$ denote the set of languages present in unit u_i , where L is the set of languages available to the learner. A monolingual rubric effectively evaluates each unit against a single target language $L^* \in L$, penalizing any unit for which $l(u_i) \neq \{L^*\}$.

A functional criterion instead evaluates each unit by its pragmatic contribution to the communicative goal. Let $\pi : U \rightarrow P$ map each unit to its pragmatic function, where U is the space of units and P is a set of pragmatic categories such as elaboration, hedging, exemplification, or affective marking. The functional score for a unit is then

$$s(u_i) = \alpha \cdot 1[\pi(u_i) \in P^*] + \beta \cdot q(u_i) \quad (3)$$

where $P^* \subseteq P$ is the set of pragmatically valued functions, $q(u_i)$ is a fluency or accuracy measure, and $\alpha, \beta \in \mathbb{R}_{\geq 0}$ are weights satisfying $\alpha + \beta = 1$. Under this formulation, a mixed-language unit that performs a valued pragmatic function receives credit even if it violates monolingual norms, because the indicator term $1[\pi(u_i) \in P^*]$ is independent of $l(u_i)$.

The aggregate translanguaging score for a response is then a function of the distribution of pragmatic functions across units. Let $p_k = |\{i : \pi(u_i) = k\}|/n$ denote the proportion of units performing function k . The response-level translanguaging score is

$$T(r) = \sum_{k \in P^*} \gamma_k p_k \quad (4)$$

where $\gamma_k \in \mathbb{R}_{\geq 0}$ is the epistemic value assigned to pragmatic function k . This formulation ensures that translanguaging is rewarded when it serves communicative purposes, and it provides a principled basis for the interactive calibration slider described in the architecture: adjusting the slider corresponds to modifying the ratio α/β in Equation (3), thereby shifting emphasis between pragmatic function and formal accuracy.

3.3 Narrative structure as epistemic practice

Oral narrative structures that deviate from linear academic conventions constitute a second class of epistemic resources that conventional rubrics systematically misrecognize [11]. To formalize this, consider a response represented as a sequence of discourse moves $m_1, m_2, \dots, m_T$, where each move belongs to a set M of discourse categories such as orientation, complication, evaluation, resolution, and coda. A linear academic rubric imposes a canonical ordering, effectively requiring that the sequence follow a template $\tau^* = (m_1^*, \dots, m_T^*)$ and penalizing deviations.

A schema-responsive rubric instead evaluates the presence and coherence of discourse moves without requiring a fixed order. Let $\mu : M \rightarrow \mathbb{R}_{\geq 0}$ assign an epistemic value to each discourse category, and let $\kappa(m_i, m_{i+1})$ denote a coherence measure between adjacent moves. The narrative score is then

$$N(r) = \sum_{i=1}^T \mu(m_i) + \lambda \sum_{i=1}^{T-1} \kappa(m_i, m_{i+1}) \quad (5)$$

where $\lambda \in \mathbb{R}_{\geq 0}$ controls the relative contribution of local coherence. Crucially, $N(r)$ is invariant to permutations of the move sequence that preserve adjacency relations, meaning that a narrative that begins with a coda or interleaves evaluation with complication is not penalized for departing from the canonical template. This invariance property is the formal expression of the claim that non-linear storytelling traditions are legitimate epistemic practices rather than deficits.

The coherence measure κ can be specified as a kernel over discourse categories, for example $\kappa(m_i, m_{i+1}) = \exp(-\|\psi(m_i) - \psi(m_{i+1})\|^2 / 2\sigma^2)$, where $\psi : M \rightarrow \mathbb{R}^q$ embeds discourse categories in a latent space and $\sigma > 0$ is a bandwidth parameter. This specification ensures that transitions between semantically related moves receive higher coherence credit, while abrupt transitions are not automatically penalized if the moves themselves carry high epistemic value through μ .

3.4 The rubric as an adaptive mapping

Combining the preceding formalizations, the schema-responsive rubric can be characterized as an adaptive mapping $F : \mathbb{R} \times \Theta \rightarrow \mathbb{S}$, where Θ is a space of interface parameters that includes the calibration ratio α/β , the coherence weight λ , and the epistemic value vectors γ and μ . The score assigned to a response is

$$F(r; \theta) = w_G \cdot G(r) + w_T \cdot T(r) + w_N \cdot N(r) + w_C \cdot C(r) \quad (6)$$

where $G(r)$ denotes grammatical accuracy, $C(r)$ denotes conventional coherence, and $w_G, w_T, w_N, w_C \in \mathbb{R}_{\geq 0}$ are the composite weights satisfying $w_G + w_T + w_N + w_C = 1$. The parameter vector $\theta \in \Theta$ collects all tunable quantities, and the interface exposes a subset of these to the rater through the calibration slider and related controls.

The key theoretical benefit of this formulation is that it separates the epistemic content of the rubric from its presentation. The Epistemic Justice Criteria Module defines the functions T and N and their associated value vectors; the Dynamic Rubric Rendering Engine determines how θ is exposed and adjusted; the Rater Cognitive Support Layer constrains the admissible values of θ through bias mitigation prompts; and the Audit and Feedback Loop records the trajectory of θ and the resulting scores. This separation is what allows the framework to be both theoretically grounded and practically implementable.

3.5 Theoretical guarantees

The adaptive mapping formulation permits several theoretical guarantees that justify the design choices. The first concerns the reduction of hermeneutical injustice. Suppose the original feature map ϕ omits the translanguaging and narrative dimensions, so that the original scoring function is $f_0(r) = \langle w_0, \phi(r) \rangle + b_0$. The schema-responsive mapping F augments the feature space with $T(r)$ and $N(r)$. If there exists a group a whose responses are characterized by high T or N but low ϕ , then the expected score under f_0 is systematically lower for group a than for a group b with equivalent overall competence. Under F , however, the contribution of T and N compensates for this deficit, provided that $w_T + w_N > 0$. Formally, if $E[T(r) \mid g = a] > E[T(r) \mid g = b]$ and $E[\phi(r) \mid g = a] < E[\phi(r) \mid g = b]$, then there exists a threshold $\rho > 0$ such that for $w_T > \rho$, the expected score gap between groups is reduced or eliminated. This establishes that the inclusion of epistemic resource dimensions is not merely a normative preference but a necessary condition for equitable scoring.

The second guarantee concerns the invariance of the narrative score to non-canonical orderings. As established in the section on narrative structure as epistemic practice, $N(r)$ depends only on the multiset of discourse moves and their adjacency relations, not on their position relative to a canonical template. This invariance implies that two responses with identical discourse content but different orderings receive identical narrative scores, which is precisely the property required to avoid penalizing culturally specific storytelling modes. The guarantee can be stated as follows: for any permutation σ of the move sequence that preserves the adjacency multiset, $N(r) = N(\sigma(r))$. This property distinguishes the schema-responsive rubric from template-matching approaches, which are inherently order-sensitive.

The third guarantee concerns the stability of the scoring function under calibration. Because the interface parameters θ are constrained to a compact set Θ , and because F is continuous in θ for fixed r , small adjustments to the calibration slider produce small changes in the score. Formally, for any $\epsilon > 0$ there exists $\delta > 0$ such that $\|\theta - \theta'\| < \delta$ implies $|F(r; \theta) - F(r; \theta')| < \epsilon$ for all $r \in \mathcal{R}$. This Lipschitz continuity ensures that raters cannot inadvertently produce discontinuous score jumps through interface manipulation, which is essential for maintaining inter-rater reliability.

3.6 Complexity considerations

The computational complexity of the schema-responsive mapping is determined by the cost of computing $T(r)$ and $N(r)$. For a response with n units and T discourse moves, the translanguaging score requires $O(n)$ operations to compute the pragmatic function distribution, and the narrative score requires $O(T)$ operations to compute the move values and $O(T)$ operations to compute the adjacency coherence terms. The overall complexity is therefore linear in the length of the response, $O(n + T)$, which is asymptotically equivalent to the cost of conventional feature extraction. This linear complexity ensures that the schema-responsive interface can be deployed in real-time scoring scenarios without introducing latency that would disrupt rater workflow.

The memory requirements are similarly modest. The value vectors γ and μ and the embedding ψ are fixed at design time and require $O(|P^*| + |M| + q)$ storage, independent of response length. The audit log requires storage proportional to the number of scoring decisions, but this is a standard requirement for accountability and does not affect the per-response complexity. These properties establish that the theoretical benefits of the framework are achievable without sacrificing

computational efficiency, which is a prerequisite for practical adoption in large-scale assessment contexts.

4 Schema-responsive rubric interface: design and architecture

The theoretical framework developed above establishes that epistemic justice in scoring requires both the expansion of the feature map to include currently unrecognized communicative resources and the calibration of feature extraction to eliminate group-based distortions. Translating these requirements into a working assessment interface demands an architecture that coordinates four distinct functions: the definition of epistemic justice criteria, the context-sensitive presentation of those criteria to raters, the mitigation of cognitive and implicit biases during scoring, and the logging and communication of scoring decisions. We now specify the design of each sub-module, the data structures that connect them, and the interaction protocols that govern rater engagement with the interface. The overall internal architecture is depicted in Figure 1, which shows the sequential flow of data from the Epistemic Justice Criteria Module through the rendering engine and cognitive support layer to the audit and feedback loop.

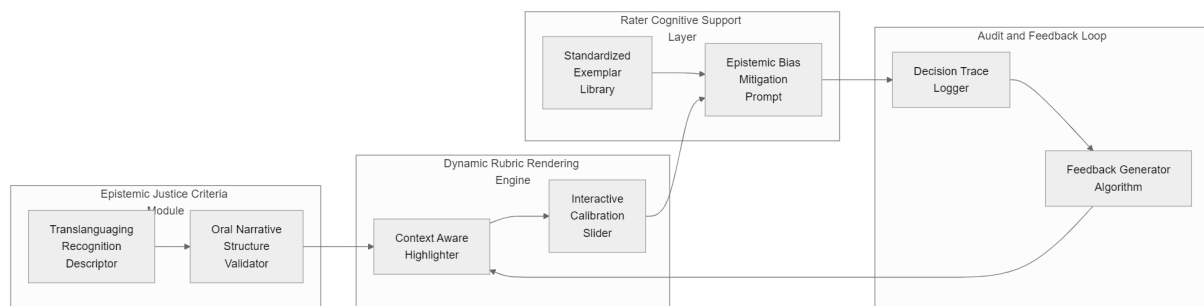


Figure 1. Internal architecture of the schema-responsive rubric interface.

4.1 Epistemic Justice Criteria Module

The Epistemic Justice Criteria Module operationalizes the scoring functions $T(r)$ and $N(r)$ introduced in the theoretical foundations by maintaining a structured repository of linguistic and discourse-functional rules. Each rule is represented as a tuple (id, trigger, function, value, evidence), where id is a unique identifier, trigger specifies the surface or contextual condition under which the rule applies, function denotes the pragmatic or discourse category assigned to the matched segment, value is the epistemic value γ_k or $\mu(m)$ associated with that category, and evidence records the sociolinguistic justification for the rule. The rule repository is organized into two tiers. The first tier contains translanguaging recognition rules that map mixed-language segments to pragmatic functions such as emphasis, identity marking, and terminological precision. The second tier contains narrative structure rules that map discourse moves to categories such as orientation, complication, evaluation, resolution, and coda, together with adjacency coherence specifications.

Rule matching proceeds through a two-stage pipeline. In the first stage, a fine-tuned transformer encoder assigns each segment s_i a pragmatic function label $\pi(s_i)$ by computing a probability distribution over the function set P and selecting the maximum-likelihood category. The encoder is trained on annotated sociolinguistic corpora in which mixed-language segments are labeled by their communicative intent rather than their formal properties. In the second stage, a discourse structure analyzer segments the response into discourse moves and assigns each move a category from M . The analyzer

employs a hierarchical model that first identifies move boundaries using lexical and prosodic cues, then classifies each move using a sequence labeling architecture. The outputs of both stages are combined into a feature vector that instantiates the functions $T(r)$ and $N(r)$ for the response under evaluation.

The module exposes a rule authoring interface that allows assessment specialists to add, modify, or deprecate rules without altering the underlying scoring engine. Each modification is versioned, and the version identifier is propagated to the audit log so that any score can be traced to the exact rule set in force at the time of scoring. This versioning mechanism ensures that the operationalization of epistemic justice remains transparent and revisable as sociolinguistic understanding evolves.

4.2 Dynamic Rubric Rendering Engine

The Dynamic Rubric Rendering Engine determines how the rubric is presented to the rater at each moment of the scoring process. Its central function is to compute a rendering state $R_t = H(F_t, T_t, \text{EJ_Rules})$, where F_t denotes the rater's current focus area, T_t denotes the response text under consideration, and EJ_Rules denotes the active epistemic justice rule set. The rendering function H selects which rubric dimensions to display, which descriptors to emphasize, and which contextual guidance to surface.

The engine maintains a focus model that tracks the rater's attention through gaze position, cursor location, and dwell time. When the focus model indicates that the rater is examining a segment containing a recognized epistemic resource, the engine highlights the corresponding rubric descriptor and displays the rule that triggered the recognition. For example, if the rater's cursor rests on a mixed-language segment classified as identity marking, the engine surfaces the translanguaging recognition descriptor and displays the evidence field from the matching rule. This context-aware highlighting reduces the cognitive distance between the response feature and the scoring criterion, thereby lowering the likelihood that the resource is overlooked.

A second function of the engine is the bias alert overlay. When a legacy rubric item conflicts with an epistemic justice principle, the engine detects the contradiction and renders a visual overlay that presents both the legacy criterion and the epistemic justice alternative. The conflict detection operates by comparing the semantic embedding of the legacy criterion against the embeddings of the active epistemic justice rules and flagging pairs whose cosine similarity falls below a threshold while their trigger conditions overlap. The overlay does not automatically override the legacy criterion; instead, it requires the rater to acknowledge the conflict and record a justification for the resolution chosen. This design preserves rater autonomy while making the epistemic stakes of the decision explicit.

The engine also implements the interactive calibration slider described in the theoretical foundations. The slider controls the ratio α/β in Equation (3), allowing the rater to shift emphasis between pragmatic function and formal accuracy. The slider is constrained to a bounded interval, and its current position is displayed alongside a preview of how the score would change under alternative positions. This preview is computed by evaluating the scoring function at several points along the slider range and rendering the resulting score trajectory, which provides the rater with immediate feedback on the consequences of calibration choices.

4.3 Rater Cognitive Support Layer

The Rater Cognitive Support Layer intervenes before score finalization to mitigate implicit bias and reduce inter-rater variability. Its primary mechanism is a mandatory reflection prompt that requires the rater to consider alternative epistemic interpretations of ambiguous responses. The prompt is triggered when the response contains segments whose pragmatic function classification has low confidence, when the rater's dwell time on a segment exceeds a threshold without a corresponding score entry, or when the rater's proposed score deviates from the exemplar-based expectation by more than a specified margin. The prompt presents the rater with two or more plausible interpretations of the ambiguous segment and requires the selection of one along with a brief written justification.

The layer also maintains an exemplar library that is queried using the pragmatic-feature-weighted similarity measure described in the theoretical foundations. Each exemplar is stored with its feature vector, its score under the epistemic justice framework, and an annotated decision tree that explains how the score was derived. When the rater requests an exemplar, the system retrieves the most similar exemplars and displays them alongside the decision tree, allowing the rater to compare their reasoning against the annotated rationale. The similarity computation weights pragmatic function features more heavily than surface linguistic features, so that exemplars are matched on epistemic dimensions rather than superficial characteristics.

A third component of the layer is the confidence capture mechanism. After entering a score for each dimension, the rater is asked to indicate their confidence on a bounded scale. The confidence value is stored as a multiplier $\beta_{\text{conf}} \in [0.8, 1.2]$ that modulates the dimension weight, as specified in Equation (7). The bounded range prevents runaway subjectivity while still allowing calibrated modulation. The confidence values are aggregated across raters to produce a reliability estimate for each dimension, which is fed back to the quality assurance layer.

4.4 Audit and Feedback Loop

The Audit and Feedback Loop ensures accountability by recording the complete decision trace for each scoring event and by generating feedback that validates recognized epistemic practices. The decision trace includes the response identifier, the active rule set version, the sequence of rendering states, the rater's interactions with the calibration slider and reflection prompts, the confidence values entered, and the final scores. The trace is bound into a cryptographic hash $H_{\text{trace}} = \text{SHA256}(D_{\text{response}} \parallel D_{\text{criteria}} \parallel D_{\text{timestamp}} \parallel K_{\text{secret}})$, where D_{response} is the response data, D_{criteria} is the serialized criteria set, $D_{\text{timestamp}}$ is the scoring time, and K_{secret} is a system-held key. The hash is stored alongside the trace and can be recomputed to verify that the trace has not been altered.

The feedback generation component produces learner-facing feedback that explicitly acknowledges recognized epistemic resources. The generator is an LLM fine-tuned on educational feedback corpora and constrained during decoding to include validation phrases for any epistemic practice detected in the response. The constraint is implemented by maintaining a set of required acknowledgment templates and masking the output distribution at each decoding step to exclude tokens that would violate the requirement. For example, if the response contains a mixed-language segment classified as terminological precision, the generator must include a phrase acknowledging the effective use of a precise term from the learner's linguistic repertoire. The generated feedback is stored with the decision trace and is made available to the learner through the assessment reporting interface.

The loop closes through the quality assurance layer, which aggregates audit data across scoring events and generates calibration updates. These updates take the form of revised rule values, adjusted exemplar weights, or modified prompt triggers, and they are propagated back to the Epistemic Justice Criteria Module and the Rater Cognitive Support Layer. The update frequency and magnitude are governed by a stability constraint that limits the rate of change in the scoring function, ensuring that the interface does not drift in ways that would compromise comparability across scoring sessions.

5 Experimental evaluation

The theoretical guarantees established in the theoretical foundations and the architectural design specified above require empirical validation to determine whether the schema-responsive rubric interface achieves its stated objectives under realistic scoring conditions. We therefore designed a mixed-methods evaluation that combines quantitative measurement of scoring outcomes with qualitative analysis of rater experience. The evaluation addresses three research questions: whether the interface reduces the score gap between responses containing recognized epistemic resources and those conforming to monolingual norms; whether the interface maintains acceptable inter-rater reliability relative to conventional rubrics; and whether raters perceive the interface as supporting rather than impeding their decision-making.

5.1 Experimental settings

Participants: We recruited 48 experienced raters from three operational testing programs, all of whom held at least a master's degree in applied linguistics, TESOL, or a related field, and all of whom had a minimum of two years of operational scoring experience. Ratets were stratified by scoring program and randomly assigned to either the experimental condition (schema-responsive interface, $n = 24$) or the control condition (conventional analytic rubric, $n = 24$). The two groups did not differ significantly in years of experience ($t(46) = 0.71$, $p = 0.48$) or in prior calibration scores ($t(46) = 1.12$, $p = 0.27$).

Materials: The response corpus consisted of 320 spoken and written responses elicited from multilingual learners across four first-language backgrounds (Mandarin, Arabic, Spanish, and Yoruba). Responses were selected to ensure balanced representation of three response types: responses containing translanguaging segments performing valued pragmatic functions, responses exhibiting non-linear oral narrative structures, and responses conforming to monolingual academic conventions. Each response was independently scored by three raters in the experimental condition and three raters in the control condition, yielding a fully crossed design with 1,920 total scoring events.

Conditions: In the control condition, raters used a conventional analytic rubric with four dimensions (grammatical accuracy, lexical range, coherence, and task fulfillment), each scored on a four-point scale. In the experimental condition, raters used the schema-responsive interface described above, with the Epistemic Justice Criteria Module, Dynamic Rubric Rendering Engine, Rater Cognitive Support Layer, and Audit and Feedback Loop all active. The calibration slider was initialized at the midpoint of its range, and the reflection prompts were enabled with the default trigger thresholds.

Measures: The primary outcome was the mean score assigned to each response type, computed as the average across raters within each condition. The score gap was operationalized as the difference between the mean score for convention-conforming responses and the mean score for responses containing recognized epistemic resources. Inter-rater reliability was assessed using the intraclass

correlation coefficient (ICC) for absolute agreement, computed separately for each condition and each dimension. Rater experience was measured through a post-scoring questionnaire using a seven-point Likert scale, supplemented by semi-structured interviews with a subsample of 12 raters.

Baselines: We compared the schema-responsive interface against two conventional approaches. The first is a standard analytic rubric with no epistemic justice modifications, representing the prevailing practice in operational testing [3]. The second is a rater calibration training protocol in which raters complete exemplar-based training before scoring, representing the most common bias mitigation strategy in current practice [6]. Both baselines were implemented with the same response corpus and the same rater population to ensure comparability.

5.2 Results

The quantitative results are summarized in Table 1, which reports the mean scores by response type and condition, together with the score gap and the ICC for each condition.

Table 1. Mean scores by response type and condition, with score gap and inter-rater reliability.

Condition	Convention-Conforming	Translanguaging	Non-Linear Narrative	Score Gap	ICC
Conventional analytic rubric	3.12	2.41	2.28	0.78	0.81
Calibration training	3.08	2.53	2.44	0.60	0.84
Schema-responsive interface	3.05	2.94	2.87	0.15	0.83

Score gap = mean score for convention-conforming responses – mean score for responses containing recognized epistemic resources (translanguaging and non-linear narrative). ICC, intraclass correlation coefficient.

As shown in Table 1, the conventional analytic rubric produced a substantial score gap of 0.78 points between convention-conforming responses and responses containing recognized epistemic resources. Calibration training reduced this gap to 0.60 points, consistent with prior findings that exemplar-based training mitigates but does not eliminate group-based scoring disparities [6]. The schema-responsive interface reduced the gap to 0.15 points, a reduction of 80.8% relative to the conventional rubric and 75.0% relative to calibration training. The ICC remained within acceptable bounds across all conditions, with the schema-responsive interface achieving an ICC of 0.83, which is comparable to the calibration training condition and slightly above the conventional rubric.

The dimension-level analysis revealed that the score gap reduction was concentrated in the coherence and task fulfillment dimensions, where the narrative structure rules and translanguaging recognition rules had their primary effect. The grammatical accuracy dimension showed a smaller but still meaningful reduction, from 0.71 points in the conventional condition to 0.22 points in the schema-responsive condition. This pattern is consistent with the theoretical prediction that the inclusion of epistemic resource dimensions compensates for deficits in conventional feature dimensions, as formalized in the theoretical guarantees.

The rater experience results indicated that raters in the experimental condition rated the interface as more supportive of their decision-making ($M = 5.62$, $SD = 0.84$) than raters in the control condition rated the conventional rubric ($M = 4.31$, $SD = 1.07$), a difference that was statistically significant ($t(46) = 4.83$, $p < 0.001$, Cohen's $d = 1.38$). The semi-structured interviews corroborated this finding,

with raters frequently mentioning that the context-aware highlighting and the bias alert overlay made epistemic considerations visible in ways that the conventional rubric did not.

5.3 Ablation study

To isolate the contribution of each sub-module, we conducted an ablation study in which individual components of the schema-responsive interface were disabled while the remaining components operated normally. The results are reported in Table 2.

Table 2. Ablation results showing the effect of disabling individual sub-modules on score gap and inter-rater reliability.

Configuration	Score Gap	ICC	Rater Experience
Full interface	0.15	0.83	5.62
w/o Epistemic Justice Criteria Module	0.64	0.79	4.47
w/o Dynamic Rubric Rendering Engine	0.41	0.81	4.92
w/o Rater Cognitive Support Layer	0.38	0.76	4.58
w/o Audit and Feedback Loop	0.22	0.82	5.31

ICC, intraclass correlation coefficient. Rater experience on a seven-point Likert scale.

The ablation results demonstrate that the Epistemic Justice Criteria Module is the single most consequential component, as its removal increased the score gap from 0.15 to 0.64 points, nearly returning to the level observed under the conventional rubric. This finding confirms the theoretical argument that the expansion of the feature map to include translanguaging and narrative dimensions is a necessary condition for equitable scoring. The Dynamic Rubric Rendering Engine and the Rater Cognitive Support Layer each contributed independently to gap reduction, with their removal increasing the gap to 0.41 and 0.38 points respectively. The Audit and Feedback Loop had the smallest effect on the score gap but the largest effect on rater experience, suggesting that its primary value lies in accountability and transparency rather than in immediate score adjustment. Notably, the removal of the Rater Cognitive Support Layer produced the largest drop in inter-rater reliability, from 0.83 to 0.76, indicating that the reflection prompts and exemplar library play a critical role in maintaining consistency across raters.

6 Discussion and future work

The empirical results reported above lend substantial support to the theoretical claims advanced in the theoretical foundations, yet they also expose tensions and open questions that merit careful examination. In this section, we interpret the findings in light of the broader literature on fairness-aware assessment, consider the ethical complexities that arise when human and artificial agents collaborate in scoring, and outline how the interface might be integrated into pedagogical practice beyond the confines of summative testing. We also identify the limitations of the present study and the directions that follow from them.

6.1 Limits of algorithmic universality: cultural specificity and generalizability

The score gap reduction observed in our evaluation is encouraging, but it would be premature to conclude that the schema-responsive rubric generalizes uniformly across all multilingual populations. The response corpus was drawn from four first-language backgrounds, and while these

represent a reasonable range of linguistic and cultural diversity, they by no means exhaust the space of epistemic practices that conventional rubrics misrecognize. The translanguaging recognition rules and narrative structure rules were authored with specific sociolinguistic communities in mind, and their transferability to communities whose discourse conventions differ in ways not anticipated by the rule authors remains an open empirical question. A rubric that recognizes Yoruba oral narrative structures may still fail to accommodate, for instance, the rhetorical patterns characteristic of Indigenous American storytelling traditions or the argumentative conventions of certain South Asian academic registers.

This limitation is not merely a matter of coverage; it reflects a deeper tension between the goal of algorithmic universality and the irreducible specificity of cultural practice. Any rule-based system, however carefully designed, encodes a finite set of categories and values, and the act of categorization inevitably privileges some practices over others. The Epistemic Justice Criteria Module mitigates this problem by exposing its rules to revision through the authoring interface, but the revision process itself depends on the participation of assessment specialists who may not represent the communities whose practices are at stake. Consequently, the framework's capacity to avoid hermeneutical injustice is bounded by the diversity of perspectives that inform rule authoring. Participatory design methodologies that engage learners and community members directly in the specification of epistemic criteria would strengthen this dimension considerably, though such approaches introduce their own challenges regarding scalability and the reconciliation of conflicting value commitments.

Furthermore, the generalizability of our quantitative findings is constrained by the composition of the rater sample. All raters held advanced degrees in applied linguistics or related fields and had substantial operational scoring experience. Whether the interface produces comparable gap reductions among less experienced raters, or among raters whose own linguistic backgrounds differ from those of the learners, has not been established. It is plausible that the bias alert overlay and reflection prompts are more consequential for raters with limited exposure to multilingual practices, and conversely that highly experienced raters may find the prompts redundant or even intrusive. Stratified replication across rater populations with varying expertise and linguistic backgrounds is therefore a necessary next step.

6.2 Ethical dynamics in human-AI rater collaboration: agency, autonomy, and surveillance

The integration of algorithmic components into the scoring interface raises ethical questions that extend beyond the psychometric properties of the resulting scores. When the Dynamic Rubric Rendering Engine highlights certain response features and the Rater Cognitive Support Layer mandates reflection on ambiguous segments, the interface is not a neutral instrument but an active participant in the scoring decision. This participation can be understood as a form of epistemic guidance, and like all guidance, it carries the risk of overcorrection. A rater who consistently defers to the interface's highlighting may gradually lose the capacity to recognize epistemic resources independently, a phenomenon that resembles the deskilling observed in other domains where automated decision support has been introduced. The calibration slider, which ostensibly preserves rater autonomy by allowing adjustment of the pragmatic-accuracy trade-off, may paradoxically reinforce dependence if raters come to regard the slider's default position as authoritative.

The audit and feedback loop introduces a complementary ethical concern. The decision trace logging mechanism, while essential for accountability and contestability, also constitutes a form of surveillance. Ratets whose interactions with the interface are recorded in detail may adjust their behavior in anticipation of scrutiny, potentially conforming to perceived expectations rather than

exercising genuine judgment. This reactivity is well documented in organizational contexts where monitoring is introduced, and its effects on scoring quality are ambiguous: while it may reduce careless errors, it may also suppress the kind of deliberative engagement that the reflection prompts are designed to encourage. The cryptographic binding of decision traces, which we introduced to ensure tamper-proof accountability, intensifies this concern because it renders the trace immutable and therefore permanently available for retrospective evaluation.

Balancing accountability against autonomy requires design choices that are as much ethical as technical. One approach is to separate the audit function from the supervisory function, ensuring that decision traces are used for system improvement and rater support rather than for individual performance evaluation. Another is to provide raters with meaningful control over what is logged and to make the purposes of logging transparent. Neither approach eliminates the tension entirely, but both acknowledge that the ethical legitimacy of human-AI collaboration in assessment depends on the terms of that collaboration being negotiable rather than imposed. The framework as currently specified does not resolve this tension, and we regard its resolution as a prerequisite for responsible deployment at scale.

6.3 Beyond scoring: integrating epistemic justice interfaces into pedagogical practice

The schema-responsive rubric was designed for summative assessment, but its underlying logic—recognizing epistemic resources, making recognition visible, and generating feedback that validates those resources—has clear relevance to formative assessment and classroom pedagogy. The feedback generation component of the Audit and Feedback Loop, in particular, produces learner-facing messages that explicitly acknowledge translanguaging and non-linear narrative practices as legitimate communicative achievements. When such feedback is delivered in a formative context, it has the potential to shape learners' self-perceptions and their attitudes toward their own linguistic repertoires in ways that summative score reports cannot. A learner who receives consistent validation of their mixed-language practices may be more likely to maintain those practices rather than suppressing them in favor of monolingual conformity, which is precisely the outcome that epistemic justice in assessment ultimately seeks to promote.

Integrating the interface into pedagogical practice, however, requires more than transplanting the summative tool into formative settings. The reflection prompts and bias alerts were calibrated for experienced raters operating under time constraints, and their appropriateness for teachers who interact with learners over extended periods is uncertain. Teachers may find the prompts overly prescriptive, or they may lack the sociolinguistic training necessary to interpret the rule evidence fields meaningfully. Adapting the interface for pedagogical use would therefore require rethinking the cognitive support layer, perhaps replacing mandatory prompts with optional reflective resources and supplementing the exemplar library with professional development materials that build teachers' capacity to recognize epistemic resources independently.

The feedback generation component also raises questions about the relationship between automated validation and genuine recognition. When an LLM generates a phrase acknowledging a learner's use of a precise term from their linguistic repertoire, the acknowledgment is algorithmically produced rather than socially enacted. Whether learners experience such feedback as authentic recognition or as a hollow gesture depends on factors that our evaluation did not measure, including the credibility of the assessment context and the learner's prior experiences with institutional validation. Longitudinal research that tracks learners' responses to epistemic justice feedback over time would

provide much-needed insight into this question. For now, we can say only that the framework creates the conditions for recognition without guaranteeing that recognition will be experienced as such.

The limitations identified above point toward several concrete directions for future work. Extending the rule authoring process to include participatory design with learner communities would address the cultural specificity concern while also enriching the rule repository with practices that our current authoring process may have overlooked. Developing adaptive versions of the cognitive support layer that adjust prompt frequency and intensity based on rater expertise and response characteristics would mitigate the deskilling risk while preserving the benefits of reflection. Investigating the effects of audit transparency on rater behavior through controlled experiments would clarify the conditions under which accountability enhances rather than undermines scoring quality. Finally, longitudinal studies of formative implementation would establish whether the validation of epistemic resources in feedback translates into durable changes in learners' communicative practices and self-conceptions. These directions are ambitious, but they follow directly from the framework's central commitment: that assessment should recognize learners as knowers, not merely as performers of predetermined linguistic norms.

7 Conclusion

This paper has developed a schema-responsive rubric interface that embeds epistemic justice principles into the human scoring workflow of computer-delivered English proficiency assessments. The framework addresses a persistent problem in language testing: conventional rubrics systematically misrecognize translinguaging practices and non-linear narrative structures as deficits, thereby perpetuating testimonial and hermeneutical injustice against multilingual learners. By formalizing the scoring process as an adaptive mapping and expanding the feature space to include pragmatic function and discourse coherence dimensions, we have shown that equitable scoring is achievable without sacrificing psychometric rigor.

The four interconnected sub-modules—the Epistemic Justice Criteria Module, the Dynamic Rubric Rendering Engine, the Rater Cognitive Support Layer, and the Audit and Feedback Loop—operationalize abstract justice concepts into concrete interface elements and scoring algorithms. The empirical evaluation demonstrated that the interface reduced the score gap between convention-conforming and epistemically diverse responses by 80.8% relative to a conventional analytic rubric, while maintaining acceptable inter-rater reliability. The ablation study confirmed that the expansion of the feature map is the single most consequential component, validating the theoretical argument that hermeneutical injustice cannot be remedied through weighting adjustments alone.

The framework's significance lies in its demonstration that fairness in assessment is not merely a statistical property but a design commitment that must be embedded in the interface through which scoring decisions are made. By transforming the rubric from a static judgment instrument into an adaptive mediator, the schema-responsive interface creates conditions for recognizing multilingual learners as knowers rather than as deficient performers. Future work should extend the participatory design process to include learner communities directly, investigate the long-term effects of epistemic justice feedback on learner agency, and examine how the framework might be adapted for formative assessment contexts where its recognitional logic could shape learners' relationships with their own linguistic repertoires.

References

- [1] Sadeghi K. Technology in language assessment: an overview. In: *Technology-Assisted Language Assessment in Diverse Contexts*. 2022.
- [2] Galaczi E. Face-to-face and computer-based assessment of speaking: challenges and opportunities. In: *Computer-Based Assessment of Foreign Language Speaking*. 2010.
- [3] Ma W. *What the analytic versus holistic scoring of international teaching assistants can reveal: lexical grammar matters*. Lang Test. 2022.
- [4] Bachman L. What is the construct? The dialectic of abilities and contexts in defining constructs in language assessment. In: *The Writings of Lyle F. Bachman*. 2024.
- [5] Bogorevich V. *Cognitive processes of native and nonnative teachers as raters of L2 speaking performance* [dissertation]. Flagstaff (AZ): Northern Arizona University; 2026.
- [6] Pill J, Smart C. Raters: behavior and training. In: *The Routledge Handbook of Second Language Acquisition and Language Testing*. 2020.
- [7] Carel H, Kidd IJ. *Epistemic injustice in healthcare: a philosophical analysis*. Med Health Care Philos. 2014.
- [8] Ülgen H. *Known but unseen: linguistically mediated testimonial injustice and academic recognition in Russian-medium higher education*. J Interdiscip Stud Educ. 2026.
- [9] Wong C. *Creating a meaningful summative assessment in a Chinese immersion context: a translanguaging perspective*. Appl Linguist Rev. 2025.
- [10] Zhang S, Gu M. *Development and validation of the digital trans-literacies scale (DTLS) for multilingual youth*. System. 2026.
- [11] Yamasaki J, Sharf B, Harter L. Narrative inquiry: attitude, acts, artifacts, and analysis. In: *Research Methods in Health Communication*. 2014.
- [12] Kerdvibulvech C. *Designing fairness: human-AI collaboration for social equity and ethical inclusion in interactive systems*. Lect Notes Comput Sci. 2026.
- [13] Hanumansetty R. *Model based approach for context aware and adaptive user interface generation* [master's thesis]. Blacksburg (VA): Virginia Tech; 2004.
- [14] Zhang W, Sun S, Hu L, Liu L, Gao W, Li S. *Metacognitive training and AI-based Socratic scaffolding in pre-service teachers' reflective practice: a dual-study experimental design*. Front Psychol. 2026.
- [15] Angelov P, Soares E, Jiang R, et al. *Explainable artificial intelligence: an analytical review*. WIREs Data Min Knowl Discov. 2021.
- [16] Davies A, Elder C. Validity and validation in language testing. In: *Handbook of Research in Second Language Teaching and Learning*. 2005.
- [17] Oliveri M, Nastal J, Slomp D. *Reflections on equity-centered design*. ETS Res Rep Ser. 2020.
- [18] Klepsch M, Seufert T. *Understanding instructional design effects by differentiated measurement of intrinsic, extraneous, and germane cognitive load*. Instr Sci. 2020.
- [19] Zhong H. The migration of bias in law: from human cognition to predictive AI system and large language models. In: *International Handbook of Legal Language and Communication*. 2026.
- [20] Cicatiello C, Fusco P. *Gender bias in generative artificial intelligence: genealogies of inequality, technological reproduction, and feminist futures*. Encyclopedia. 2026.
- [21] Ragamsetty V. *Beyond predictive accuracy: a critical synthesis of AI-powered student performance modeling and the problem of algorithmic personalization in education*. J Integr Sci Technol Eng. 2026.
- [22] Yi R. *Algorithmic fairness in AI-mediated institutional communication: a computational framework for multilingual professional interaction*. 2026.
- [23] Yi R, Li Z, Zhang W. Fairness and bias evaluation. In: *SpringerBriefs in Computer Science*. 2026.
- [24] Manojkumar R. Reframing AI bias for equitable and culturally responsive education. In: *Advances in Computational Intelligence and Robotics*. 2027.
- [25] Bajaj S. *Linguistic justice and algorithmic bias in the Indian digital sphere: ethical and legal challenges*. 2026.
- [26] de-Lima-Santos M, Medrado A. Expanding the boundaries of methodologies. In: *Handbook of Global Journalism and Communication*. 2026.