International Journal of Learning, Teaching and Educational Research
Vol. 25, No. 4, pp. 916-935, April 2026
https://doi.org/10.26803/ijlter.25.4.41
Received Nov 15, 2025; Revised Jan 27, 2026; Accepted Apr 6, 2026

Reconceptualising Fair Assessment in English-Medium Instruction: Content–Language Separation as a Validity Principle in Higher Education

Randip Kaur Valdev Singh*, Harwati Hashim and Khairul Azhar Jamaludin
Universiti Kebangsaan Malaysia (UKM)
Bangi, Selangor, Malaysia

Abstract. This narrative review promotes fair assessment in English-medium instruction in higher education, in order to align with Sustainable Development Goals 4 and 10 and to ensure that grades reflect disciplinary learning rather than English fluency. This narrative review draws on searches of Scopus, Web of Science, and ERIC to access studies published from 2022 to 2025. It synthesizes evidence and highlights recurring trends in the way assessment design, scoring, and moderation can keep content and language on analytically separate strands across three assessment families: written products, oral or performance tasks, and tests or portfolios. Guided by validity and argument-based and socio-cognitive perspectives, the review traces how dense reading passages, speeded conditions, monolingual orientations, translanguaging constraints, and rater inconsistency can pull scores toward fluency or test-wiseness instead of disciplinary knowledge. Findings indicate that fairness improves when the construct is stated explicitly at design, and when content and language are separated in scoring, linguistic load is managed through readability and lexical coverage checks, and brief calibration and moderation routines stabilize judgment. These routines can be embedded in ordinary timetables, staffing, and resources, and support more valid, interpretable, and equitable decisions. Overall, explicit content–language separation, from task design through scoring and moderation, offers a scalable organizing principle for fair assessment in English-medium instruction, and the review sets out practical routines that programs and higher education institutions can enact with transparency.

Keywords: English-medium instruction; fair assessment; content–language separation; higher education; Sustainable Development Goals

Citation:
Singh, R. K. V., Hashim, H., & Jamaludin, K. A. (2026). Reconceptualising Fair Assessment in English-Medium Instruction: Content–Language Separation as a Validity Principle in Higher Education. *International Journal of Learning, Teaching and Educational Research*, *25*(4), 916–935. https://doi.org/10.26803/ijlter.25.4.41

*Corresponding author: Randip Kaur Valdev Singh; *randipvaldev@gmail.com*

1. Introduction

Higher education has expanded English-medium instruction (EMI), and assessment is the point at which opportunity and risk converge. Students are expected to demonstrate disciplinary understanding in a language that may not be their strongest, which creates tension between what students know and how fluently they can express it. When linguistic load or time pressure is high, scores can drift toward signaling control of English rather than disciplinary attainment. Dimova and Kling (2022) show that apparently small task-design choices, such as option order and timing, introduce variability unrelated to mastery, while Karanfil and Neufeld (2020) report similarly that formal task features can distort score meaning. Malmström et al. (2025) found that dense vocabulary and complex syntax depress performance in reading-heavy tasks because students redirect cognitive effort from reasoning about content to decoding language.

These pressures intensify in diverse cohorts where language proficiency, prior schooling, and disciplinary background vary widely, and Aizawa et al. (2025) report that prior exposure to EMI can mitigate or amplify the impact of linguistic control on academic outcomes. Without explicit controls at assessment design and scoring, disciplinary constructs blur and rater judgment becomes inconsistent and undermines equity goals. From a validity perspective, this drift reflects construct-irrelevant variance and weakens the interpretive argument for score meaning in EMI settings (Messick, 1989; Kane, 2013), particularly when task demands and language load are not aligned with the intended construct (Weir, 2005).

This review is guided by three complementary theoretical perspectives that frame fairness as a matter of validity in EMI assessment. Messick's (1989) unified validity theory positions fairness as integral to valid score interpretation, Kane's (2013) argument-based approach emphasizes the need to justify links between performance and decisions, and Weir's (2005) socio-cognitive framework highlights how task conditions and rating processes shape opportunities to demonstrate knowledge. Together, these perspectives foreground the central problem addressed in this review: construct-irrelevant variance arising from language demands in EMI assessment.

In parallel, quality assurance and accreditation expectations emphasize transparent and defensible assessment decisions, which makes fairness and validity in EMI assessment a policy-relevant concern. International higher education frameworks, including outcomes-based accreditation systems and quality assurance standards, increasingly require that assessment practices demonstrate validity, reliability, and equity, particularly in linguistically diverse contexts. This strengthens the need for assessment designs in EMI that clearly distinguish disciplinary achievement from language proficiency.

Accordingly, this narrative review examines how assessment can be designed and enacted so that grades reflect disciplinary learning rather than English fluency, by focusing on the principle of content–language separation across assessment processes.Across assessment formats, the literature documents parallel risks. Gronchi (2024) explains that panel conduct can shift outcomes in vivas or , and

Iskandarova (2024) reports that variability in rater behavior can tilt decisions toward rewarding delivery rather than conceptual accuracy. In reading-heavy tests, Holzknecht et al. (2022) and Li et al. (2024) found that inadequate lexical coverage suppresses performance even when underlying understanding is intact. Workable safeguards exist but are unevenly implemented; these safeguards include construct-explicit task briefs, dual-strand rubrics that separate content from language, annotated exemplars used for brief norming, readability checks with basic item diagnostics, and moderation routines that stabilize interpretation over time.

Against this backdrop, this narrative review makes a constructive case for fair decision-making in EMI in higher education assessment that involves separating content evidence from language evidence. Its objective is to synthesize current evidence on fair assessment in EMI and to distill practical, scalable routines for content–language separation across written products, oral or performance tasks, and tests or portfolios, by identifying high-leverage practices that programs can adopt under ordinary conditions, from task design through to scoring and moderation.

The remainder of this review is structured as follows. The next section outlines the theoretical and conceptual foundations underpinning fair EMI assessment, followed by the methodology used to conduct the narrative review. The results section presents a thematic synthesis across assessment families, which is then interpreted in the discussion through validity and socio-cognitive perspectives. The review concludes with implications for practice, policy, and future research in higher education EMI contexts.

2. Literature Review

2.1 Validity and Socio-Cognitive Approaches to Fair EMI Assessment

A validity-first stance anchors fair assessment in EMI. This review draws on two complementary perspectives that keep the disciplinary construct visible in higher education. From a validity perspective, Messick (1989) offers a unified account in which fairness is integral to validity rather than being an optional add-on: Score meaning must represent the intended construct, avoid underrepresentation, and limit variance from construct-irrelevant features.

In EMI, achieving this requires distinguishing content criteria, such as conceptual accuracy, disciplinary reasoning, and warranted conclusions, from language criteria, such as organization, register, and intelligibility, so that scores primarily reflect the construct of interest. Kane's (2013) argument-based approach extends this view by requiring that the chain from task to score to interpretation and use is articulated and supported by evidence. A recent validity perspective reinforces this argument-based view by emphasizing how validity evidence must align with intended interpretations and uses in educational assessment (Chapelle, 2021). In this framework, content–language separation becomes an explicit design decision that clarifies which aspects of performance count as evidence for content claims and which relate to language, and on what basis decisions in higher education can be justified. These perspectives can be distinguished in terms of their primary

focus, with validity-based approaches emphasizing score interpretation and justification, while socio-cognitive approaches foreground task conditions and performance processes.

Weir's (2005) socio-cognitive framework complements these accounts by examining how task input, processing demands, response or interactional formats, test conditions, and rating procedures jointly shape opportunities to demonstrate knowledge. When it is applied to EMI, fairness improves when the assessment purpose is declared as content focused or language focused, when input and processing demands align with that purpose, for example, through control of reading load and clarity of instructions, and when interactional formats reduce interlocutor effects in oral or performance tasks.

On the scoring side, strand-specific criteria, short annotated exemplars, and brief calibration activities help keep rater attention on the intended construct rather than on global impressions of fluency. Across written products, oral or performance tasks, and tests or portfolios, these design and procedural choices preserve the evidential link between performance and construct. Taken together, the validity and socio-cognitive perspectives advanced by Messick (1989), Kane (2013), and Weir (2005) clarify what should be inferred, how assessment conditions can obscure it, and why explicit content–language separation, operationalized through aligned inputs, processing demands, formats, and rating, is central to fair interpretation in higher education.

Across the EMI assessment literature synthesized in this review, qualitative and small-scale studies are more common than quantitative or mixed-methods designs, and reporting depth varies across assessment families. Figure 1 presents a conceptual model of content–language separation as an organizing principle for fair EMI assessment.

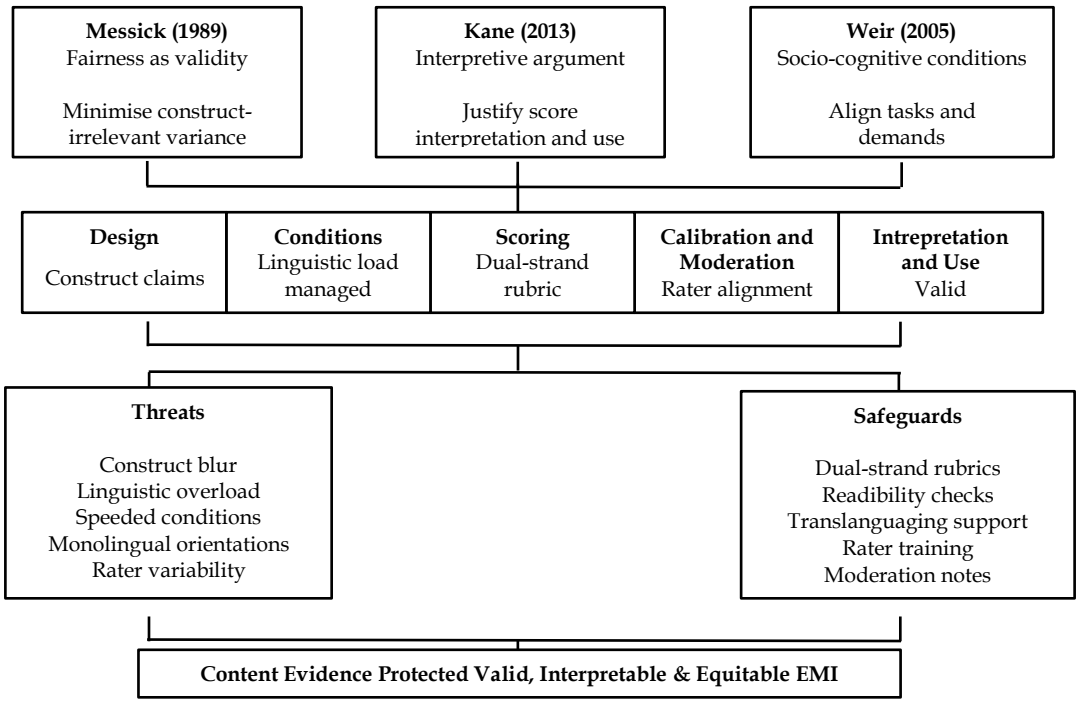

Figure 1: Content–language separation as an organizing principle for fair EMI assessment

As shown in Figure 1, fair EMI assessment depends on maintaining alignment between construct definition, task conditions, scoring processes, and interpretation. Together, validity-based and socio-cognitive perspectives clarify both what should be inferred from performance and the conditions under which learners can demonstrate disciplinary knowledge, thereby reinforcing the need for explicit content–language separation in EMI assessment.

2.2 Design and Moderation as Mechanisms for Content–Language Separation in EMI Assessment

At the design stage, fair assessment in EMI starts by stating the disciplinary claims tasks are intended to elicit and keeping content and language on clearly separated strands so that the construct is visible to students and raters. Sato (2023) shows that, when briefs and rubrics signal this separation, students plan to produce disciplinary evidence and markers read for ideas rather than surface polish.

Regarding writing, Graham and Eslami (2020) illustrate how strand-labeled criteria recenter concept accuracy, methodological reasoning, and warranted conclusions, and counter fluent prose that can masquerade as mastery. Evidence-centered design (Mislevy et al., 2002) provides a template by specifying the claim, the observations that support it, and the rules for interpreting those observations, so that language functions primarily as a medium for access rather than the target of judgment. Building on this approach, Morton (2022) and Morton and Nashaat-Sobhy (2023) explain how cognitive discourse functions provide a shared vocabulary for the content strand; mapping prompts moves such as explaining,

arguing, and evaluating, and concluding turns the construct into observable action while keeping language expectations parallel but analytically separate.

Operationalizing these principles requires shared tools and routines. Morton et al. (2021) report that annotated exemplars and co-constructed rubrics strengthen stakeholders' grasp of criteria, while Ait-Hroch et al. (2025) state that constructive alignment across objectives, tasks, and assessment reinforces rubric clarity and consistency. Gonsalves (2023) argues that clear dual-strand rubrics can reduce linguistic barriers while building assessment literacy for teachers and students. Inclusion can be supported without redefining the construct when multilingual resources function as scaffolds rather than additional criteria. Gülle and Bayyurt (2024) describe bounded translanguaging in which concise first-language definitions, technical terms, or brief repairs are permitted as access support but excluded from the language score, and Xin and Yap (2025) and Choi et al. (2020) explain that multilingual repertoires can be specified in task briefs and support materials. Inbar-Lourie (2022) and Lo and Leung (2022) recommend formative checkpoints, before summative judgment, that keep strands distinct and provide targeted feedback on disciplinary reasoning.

When evidence is oral or performance-based, design and conduct become central to fairness. Pearce and Chiavaroli (2020) and Wang (2021) recommend short, structured prompts and concise, concept-focused question-and-answer sequences to stabilize the responses elicited in vivas, defenses, and similar tasks from candidate to candidate. In applied laboratory and clinical settings, according to Hou et al. (2024), procedural checklists and consistent examiner talk foreground accuracy and problem-solving, while Bozbıyık et al. (2024) indicate that bounded translanguaging can make procedural knowledge more visible when scope and recording rules are clearly specified. To secure consistency across raters and cohorts, moderation routines are necessary. Middleton et al. (2024) found that brief, consensus-oriented calibration at the start of a marking cycle narrows rater spread and recentres the content strand; Karnas-Haines (2021) found that calibration and social moderation built shared understandings of outcomes and criteria; and Chambers et al. (2024) describe comparative judgment as a technology-supported route to reliable moderation that also supports exemplar banks.

Beyond individual tasks, routine technical checks and transparent policy help sustain fairness at program level. Macaro et al. (2018) advise stating test purpose (primarily content or language) in candidate materials and marker guidance to avoid construct blur. Before delivery, Karanfil and Neufeld (2020), Lumban Raja (2020), and Park and Wright (2023) recommend reviewing readability and lexical coverage and checking option order for unintended difficulty shifts; after administration, Syukur and Nurlaily (2025) and Hidayati and Andriyanti (2025) show that analyses of item difficulty, discrimination, and distractor efficiency guide revision and inform instructional responses. Macaro et al. (2018) caution that rigid English-only expectations inflate construct-irrelevant variance and suppress engagement, whereas transparent access standards and readable test materials support both fairness and participation. Correspondingly, Pearson

(2021), Wudthayagorn (2025), Owen and Senel (2025), and Inbar-Lourie (2022) argue that program guidance that specifies dual rubrics for written and oral or performance tasks, documented standard setting, readability expectations, and clear procedures for bounded translanguaging makes decisions auditable and portable across cohorts and campuses. The through-line is that design makes the construct explicit, procedures keep it in view during scoring and moderation, and policy sustains both, thereby fair assessment in EMI becomes routine practice in higher education, rather than an exception.

2.3 Sources of Construct-Irrelevant Variance in EMI Assessment

Fair assessment in EMI depends on how evidence of learning is elicited, interpreted, and judged, yet these processes often tilt toward linguistic polish rather than disciplinary understanding. Sato (2023) explains that polished prose can be misread as mastery, and Graham and Eslami (2020) found that hesitant English can conceal sound disciplinary reasoning, which indicates that raters may privilege linguistic surface over conceptual depth. Construct blur and under-specified constructs weaken the link between evidence and inference further.

Şahan and Şahan (2022) report that language control leaked into content scores in Turkish engineering programs. Similarly, Gronchi (2024) documents how intended distinctions between assessing content and language can slip back into impressionistic marking. In a related vein, Lin (2023) observes that ambiguous construct boundaries leave raters without clear definitions of disciplinary evidence. Together these studies highlight the central challenge, that of maintaining transparent content–language separation so that scores reflect disciplinary knowledge rather than English fluency.

Monolingual orientations that sideline multilingual scaffolds can also narrow fairness by hiding disciplinary reasoning. Lin (2023) found that excluding translanguaging restricts cognitive access, whereas Gülle and Bayyurt (2024) report that bounded translanguaging (e.g., a brief first-language definition followed by a second-language paraphrase) can reduce linguistic load without compromising the construct. Murray (2022), similarly, demonstrates that integrated language-support models strengthen interpretation across EMI classrooms. Within a balanced content–language framework, such multilingual mediation operates as access support rather than a competing construct.

Uneven EMI assessment literacy of academic staff compounds these issues in maintaining content–language separation. Inbar-Lourie (2022) found that staff without a practical grammar for separating content and language tend to default to global impressions, and Graham and Eslami (2020) explain that feedback often targets form more than reasoning. Crain and Bailey (2022) state that criterion-specific comments promote deeper revisions. These findings underline the need for sustained professional learning to support content–language separation in everyday assessment practice.

Calibration and role clarity are critical for stabilizing judgment, yet they are often weakly implemented. According to Watari et al. (2022) short calibration using shared exemplars recentres scoring on disciplinary evidence, and Şahan and Şahan (2022) recommend locally tuned standards that reflect disciplinary norms. Lo and Leung (2022) report that blurred responsibilities over who judges content or language invite inconsistency. Panel dynamics and examiner talk in vivas and performance tasks can distort the quantity and quality of elicited evidence further. Mayahi and Jalilifar (2022) indicate that questioning patterns and interactional cues influence how much disciplinary reasoning candidates can display, while Cade and Meuller (2024) found that unstructured examiner sequences obscure procedural traces and weaken interpretation. Clarifying examiner scripts, distinguishing interactional scaffolding from conceptual evaluation, and defining scoring roles are, therefore, important for preserving content–language separation in oral and performance-based assessments.

For reading-dependent formats, heavy reading load and linguistic complexity can overload working memory and reduce access to disciplinary reasoning. Li et al. (2024) demonstrate that dense terminology and long multi-clause sentences depress comprehension, even for proficient readers, and de-la-Peña and Luque-Rojas (2021) report steep drops from literal to inferential and critical comprehension as linguistic density increases. Stoeckel et al. (2021) found that systematic readability checks can sustain cognitive demand without imposing additional linguistic penalty. Under testing pressure, speeded conditions and option or sequence effects can, furthermore, displace reasoning with test-wiseness. Ardoin et al. (2024) explain that time pressure encourages surface strategies, and Al Fraidan (2024), Karanfil and Neufeld (2020), and Asquith (2022) show that guessing, elimination, and reliance on option position increase as vocabulary load rises. These findings indicate that regulating timing and lexical coverage is essential if success in EMI assessments is to be driven by content rather than fluency or test-taking strategies.

Underused controls for readability, explicit test purpose, and basic diagnostics leave interpretations unstable. Stoeckel et al. (2021) report that routine readability and lexical coverage checks increase access without lowering cognitive difficulty, and Yousefpoori-Naeim et al. (2023) explain that declaring whether a test targets content or language reduces construct blur. Motavas and Mahmood (2025) link regular analysis of item difficulty, discrimination, and distractor efficiency to tighter alignment with disciplinary knowledge. After scoring, feedback and moderation, practices can still drift toward surface correctness. Sato (2023) notes that comments often fixate on error correction, while Watari et al. (2022) demonstrate that mid-cycle calibration with exemplars narrows rater variance, and Crain and Bailey (2022) confirm that detailed, criterion-based feedback deepens learning. Embedding diagnostics, moderation loops, and reflective commentary into routine practice helps maintain fair assessment integrity in EMI.

At institutional and system levels, conditions can either sustain or erode fairness. According to Macaro et al. (2018), English-only mandates heighten language-led variance, and Inbar-Lourie (2022) observed that throughput pressures and

blurred responsibilities can erode stabilizing routines such as readability checks, item vetting, and light moderation. White and Ronfeldt (2024) found that concise marker notes and explicit moderation roles reduce inconsistency. Gülle and Bayyurt (2024) confirm that bounded translanguaging can maintain validity, Ojochegbe (2024) advocates for culturally responsive assessment that recognizes diverse repertoires of meaning-making, and O'Donovan et al. (2024) emphasize that social moderation and calibration dialogues help align standards across institutions. Taken together, these studies suggest that equity architecture and system design determine whether fairness persists across programs and cohorts. Advancing fair assessment in EMI thus requires transparent and sustainable systems in which content–language separation is explicitly embedded and grades function as valid indicators of disciplinary knowledge and capability rather than linguistic advantage.

3. Methodology

This review synthesizes work on fair assessment in EMI by paying explicit attention to content–language separation in higher education. Narrative reviews are suitable for theory building and field mapping because they integrate diverse study types into a coherent argument about what matters, and why (Baumeister & Leary, 1997; Green et al., 2006).

The review focused on EMI assessment in higher education. Searches were conducted in major education and applied linguistics databases, and supplemented by backward and forward citation tracking from key articles and relevant institutional or professional guidance documents. Search terms, data sources, and inclusion criteria are summarized in Table 1. Study selection was iterative: Titles and abstracts were screened for relevance to assessment, fairness or treatment of content and language in EMI; full texts were then examined to identify contributions to organizing themes, including task and rubric design, scoring and feedback routines, moderation and calibration, technical test quality, and equity-oriented access supports. Studies were selected according to their relevance to the review focus and contribution to the identified themes. The review prioritized contemporary EMI and assessment research, and incorporated earlier foundational work if it had remained theoretically or methodologically influential.

Analysis was interpretive. Data management and analysis were conducted manually without the use of specialized software. Evidence was grouped around three assessment families (written products, oral or performance tasks, and tests or portfolios) and mechanisms consequential for fairness, including dual-strand rubrics, calibration practices, readability checks, and translanguaging arrangements. Categories were refined as new studies were incorporated, and studies were foregrounded when they offered conceptual leverage or detailed accounts of assessment routines. The aim was not exhaustive coverage but a structured, conceptually grounded synthesis of how fair assessment in EMI can be advanced through explicit content–language separation. Because this was a narrative review, no formal quality appraisal checklist was applied; however,

priority was given to peer-reviewed and methodologically transparent studies. This approach may be subject to selection bias and limited systematicity.

Table 1: Search and inclusion criteria

Category	Details
Keywords	"English-medium instruction"; "EMI assessment"; "fair assessment"; "equity"; "content–language separation"; "dual-strand rubric"; "moderation"; "calibration"; "readability"; "translanguaging"; "higher education"
Data sources	Scopus; Web of Science; ERIC; selected journals for language assessment, applied linguistics, and higher education; backward and forward citation tracking; institutional and professional guidance documents
Inclusion criteria	Higher education context; EMI; focus on assessment of disciplinary learning, language or both; explicit relevance to fairness, validity, equity, moderation or access supports; empirical or conceptual scholarship and institutional or professional guidance; published in English

4. Results

Across the reviewed studies, a recurring finding is that three assessment families emerge: written products, oral or performance tasks, and tests or portfolios. Fair assessment is most consistently supported when evidence of disciplinary learning is distinguished from evidence about students' English use. When this distinction is blurred, fluent linguistic performance is easily misinterpreted as stronger attainment, and the validity and defensibility of assessment decisions are weakened.

For written products, fair assessment depends on defining the disciplinary construct first and then treating language as a separate, explicitly described strand. Across a majority of reviewed studies, markers are more consistent when they attend to disciplinary ideas, methodological appropriateness, and warranted claims as content, while judging language in terms of intelligibility and suitable register rather than near-native accuracy. Dual-strand rubrics, genre-informed descriptors, and brief calibration around exemplars help academic staff maintain this distinction in practice and reduce disputes about whether writing quality overshadows substantive knowledge.

In oral and performance assessments, including presentations, vivas, and laboratory or objective structured clinical examination (OSCE) type stations, findings suggest that fairness is influenced by task design and panel routines. Unstructured questioning and uncalibrated panels invite construct-irrelevant variation linked to raters' language expectations and interactional preferences. When prompts are aligned with clearly specified content targets, panels are pre-briefed, and short calibration activities are used, and raters more reliably separate clarity of delivery from the quality of disciplinary reasoning. Allowing bounded translanguaging as an access support while keeping the scored construct in the content domain reduces the risk that language barriers mask subject-matter understanding further.

For tests and portfolios, the main fairness concern is alignment between the declared purpose of the assessment and its linguistic demands. Studies of EMI examinations show that, when content-focused tests control reading load and lexical complexity and are supported by basic item analysis and standard-setting procedures, score interpretations are more robust and defensible. Portfolio studies, similarly, report fairer outcomes when samples are moderated against criteria that distinguish progression in disciplinary learning from improvements in surface polish alone. These routines are associated with reduced construct-irrelevant variance arising from heavy reading demands, technical flaws in items, and portfolio drift away from the intended learning outcomes.

Taken together, the three assessment families offer complementary routes for operationalizing content–language separation at the stages of design, scoring, and moderation. When content and language strands are specified and weighted separately, written products make disciplinary thinking visible. Oral and performance tasks become more equitable when they are choreographed to elicit comparable disciplinary evidence across students and panels, and when language is treated primarily as a vehicle for access rather than as the construct itself. Tests and portfolios remain interpretable when linguistic demand is deliberately controlled and monitored through routine technical checks. Table 2 summarizes, for each assessment family, the typical formats, the main forms of content and language evidence, the principal threats to fairness, and the key routines used to strengthen content–language separation in EMI contexts.

Table 2: Fair assessment in EMI by assessment family: content–language separation

Assessment family	Written products	Oral or performance	Tests or portfolios
Typical forms	Essays; discipline-specific writing; project reports; portfolios	Presentations; vivas or defenses; laboratory or OSCE stations	Content examinations; multiple choice questions or reading-heavy tests; cumulative or multimodal portfolios
Content evidence	Concept accuracy; disciplinary reasoning, methods and analysis; warranted conclusions; genre or cognitive discourse function moves	Concept accuracy; justification of choices; transfer to novel cases; procedural correctness and safety; interpretation of results	Correct solutions; application of concepts; integrated evidence of progression
Language evidence	Organization and cohesion; academic or disciplinary register; clarity appropriate to level	Intelligible delivery; accurate terminology; interactional clarity	Reading and lexical demands; reflective or explanatory language in artifacts
Key risks to fairness	Construct conflation; language eclipsing content; inconsistent criteria; assessor	Rater variability; panel dynamics; uneven probing; language barriers	Linguistic load and readability masking content; option and sequence effects; time

Assessment family	Written products	Oral or performance	Tests or portfolios
	literacy gaps; portfolio polish drift	masking content; station-to-station inconsistency	pressure; summative dominance; portfolio drift
Practices that improve fairness	Dual-strand rubric for content and language; genre and CDF anchoring; brief assessor guidance; benchmarked exemplars; light calibration; formative checkpoints with targeted feedback on the content strand	Structured prompts; short concept-focused question-and-answer; pre-brief panel calibration; benchmark scripts or videos; procedural checklists; targeted double marking; bounded translanguaging as an access accommodation (not a scored trait)	Declare test purpose as content or language; apply readability and lexical control; run item diagnostics for difficulty, discrimination, and distractor efficiency; vary option order; conduct standard setting; sample portfolios periodically; maintain transparent dual criteria
References	Graham and Eslami (2020) Inbar-Lourie (2022) Lo and Leung (2022) Morton (2022) Morton and Nashaat-Sobhy (2023) Sato (2023)	Gronchi (2024) Gülle and Bayyurt (2024) Iskandarova (2024) Jalilifar and Mayahi (2022) White and Ronfeldt (2024)	Al Fraidan (2024) Dimova and Kling (2022) Holzknecht et al. (2022) Inbar-Lourie (2022) Li et al. (2024) Malmström et al. (2025) Otto and Estrada Chichón (2021)

5. Discussion

This discussion interprets the synthesis through three complementary lenses to explain why fair assessment in EMI depends on explicit separation between disciplinary content and language use. Messick (1989) provides the validity foundation, Kane (2013) supplies an interpretive argument structure that links observations to decisions, and Weir (2005) clarifies how task conditions and rating procedures shape opportunities to display knowledge. Together, these frameworks locate fairness within validity and show how it can be sustained in routine assessment practice.

From a validity perspective, Messick (1989) treats fairness as integral to a unified account of validity: Score meaning must represent the intended construct, avoid underrepresentation, and minimise variance arising from construct-irrelevant features such as language form when knowledge is the target. In EMI, this implies distinguishing content criteria, such as conceptual precision and disciplinary reasoning, from language criteria, such as organization, register, and intelligibility. When these strands are explicitly separated and made visible in

scoring, variance attributable to language form is contained and interpretations are better aligned with the disciplinary construct. Content–language separation, therefore, functions as a mechanism for reducing construct-irrelevant variance in multilingual higher education settings.

Within Kane's (2013) argument-based framework, fairness is examined through the chain of inferences that connect task performance to score interpretation and subsequent decisions. Each link must be articulated and supported by evidence. Content–language separation clarifies what counts as evidence for content claims and what counts as evidence for language claims; when tasks, scoring scales, and moderation routines specify these links, the interpretive argument becomes more coherent and defensible. Reducing noise from language form, when language is not the primary construct, strengthens warrants for high-stakes decisions such as progression and certification and supports transparency and accountability in EMI assessment.

Weir's (2005) socio-cognitive model adds a focus on how task input, processing demands, administration conditions, interactional format, and rating procedures together determine opportunities to demonstrate knowledge. Fairness improves when the declared purpose of an assessment, whether content focused or language focused, is aligned with its input characteristics and cognitive load, and when rating procedures concentrate on the intended strand rather than on global impressions of fluency. The routines identified in this review, including structured prompts, basic readability control, brief calibration using exemplars, and targeted double marking, can be understood as socio-cognitive controls that keep the construct visible under realistic teaching and learning conditions in EMI courses.

Taken together, Messick (1989), Kane (2013), and Weir (2005) converge on a shared requirement consistent with the empirical synthesis: Assessment design should state the construct explicitly and separate content and language strands where appropriate; the interpretive argument should trace how observations lead to defensible decisions, with clear warrants for handling language-related evidence; and the socio-cognitive layer should ensure that tasks and rating procedures allow students a fair chance to display disciplinary knowledge without undue interference from language demands when language is not the main object of assessment. In this light, dual-strand rubrics, calibration around annotated exemplars, readability and lexical control in test items, structured prompts with concept-focused questioning, procedural checklists, and transparent standards are coordinated means of stabilizing interpretations in EMI assessment.

At the same time, language is not always ancillary. In some EMI contexts, disciplinary communication and academic literacy are intended learning outcomes. In such cases, fairness does not mean removing language from the construct, but specifying its role explicitly and proportionately. Content–language separation remains relevant because it distinguishes when language is assessed as part of disciplinary competence and when it is treated as an access condition.

The reviewed literature is unevenly distributed across regions, institutions, and disciplines, with stronger representation from well-resourced and predominantly Western institutional contexts and small-scale studies. Many routines recommended in the literature are, therefore, theoretically coherent and practically promising, but not yet widely tested across diverse EMI ecologies. In this respect, the present synthesis extends related reviews that document recurring EMI assessment challenges by organizing risks and safeguards across assessment families and stages and by specifying content–language separation as the operational link between fairness arguments and everyday assessment practice.

This aligns with prior reviews that similarly identify persistent tensions between language and content assessment in EMI contexts. The implications are particularly salient for Global South and non-Western EMI contexts, where linguistic diversity and resource constraints may intensify fairness risks while also increasing the value of low-burden, transparent routines such as dual-strand rubrics and brief calibration, as EMI implementation often occurs under differing institutional, linguistic, and resource conditions.

Overall, fair assessment in English-medium higher education rests on systematic design and validation rather than post hoc correction of perceived bias. Explicit attention to how content and language are represented, interpreted, and operationalized aligns with Messick's validity foundations, Kane's interpretive argument, and Weir's socio-cognitive conditions. When implemented consistently across written products, oral or performance tasks, and tests or portfolios, these principles help ensure that grades and other high-stakes decisions reflect disciplinary attainment rather than linguistic ease.

6. Limitations and Recommendations

This review focuses on three assessment families used in higher education: written products, oral or performance tasks, and tests or portfolios. The scope supports transferability across disciplinary traditions and language-policy contexts. Three limitations apply. First, empirical coverage is uneven across regions and institutional types, therefore some practices may be over-represented while others remain under-documented.

Second, institutions differ in how English functions as a medium of instruction, a target of learning, or a communicative resource, which shapes what counts as acceptable access support and how assessors interpret evidence. Third, reporting is inconsistent for technical aspects such as readability checks, item statistics, and indices of rater agreement, which constrains direct comparison. These features are treated as bounded conditions rather than defects, and the synthesis, therefore, concentrates on adaptable routines such as content–language separation and dual-strand design with light calibration.

Recommendations follow the three assessment families and address both practice and reporting. For written products, future work should broaden disciplinary coverage and report how dual-strand rubrics are implemented, how exemplars

are annotated and used, and how often calibration or second marking is conducted. For oral or performance assessments, studies should document panel preparation, questioning structure and focus, procedural checklists, and the scope of translanguaging arrangements permitted as access support. For tests and portfolios, studies should state whether an assessment is primarily content focused or language focused, describe readability controls, and report basic item- and portfolio-level statistics relevant to fairness.

Across all three families, more consistent attention to the people and systems around assessment would strengthen the evidence base. Future studies should describe examiner experience, assessment literacy, and typical scoring styles, and include perspectives from students, course leaders, moderation committees, and academic support units to clarify implementation climates and the extent to which routines depend on local assessment cultures. Longer-term follow-up is needed to examine whether clearer construct specification, regular but manageable calibration, and principled access supports sustain fair assessment over time in EMI programs, and whether these practices remain feasible under real workload and legacy examination constraints. Overall, more systematic reporting and pragmatic adoption of these routines are likely to consolidate content–language separation as an organizing principle for fair assessment in higher education.

7. Implications

The implications of this review center on how content–language separation can be translated into everyday assessment routines in English-medium higher education. For practice, clearer signaling of constructs and dual-strand feedback can help students distinguish disciplinary evidence from language control, while still supporting students' academic writing and speaking development. This clarity can support retention and progression in EMI programs, particularly for multilingual students.

For academic staff, ongoing professional development that treats content and language as related but distinct strands, supported by reusable exemplars and concise calibration notes, can promote more consistent judgment within and across courses. At an institutional level, fairness by design is more likely to take root when access supports are specified in program documentation, aligned with intended outcomes, and periodically reviewed so that accommodations increase access without redefining the construct. At policy level, guidance that requires declared task purpose (content focused or language focused), sets expectations for readability, mandates transparent moderation records, and provides a basis for quality assurance in which fair assessment is inspectable rather than assumed.

For research, the framework clarifies what should be measured and how comparisons can be made in EMI assessment. Studies that report readability methods and thresholds, core item-analysis statistics, and indices of rater agreement enable clearer distinctions between variance attributable to disciplinary learning and variance attributable to language demands. Longitudinal designs can test whether gains associated with calibration, explicit standards, and principled access supports are sustained over time and across

cohorts. Cross-site work that documents fidelity indicators, assessor preparation, and local assessment cultures can explain why content–language separation yields larger fairness gains in some settings than others. Taken together, aligning pedagogy, moderation practices, institutional policy, and research methods around explicit attention to how content and language are treated provides a coherent path toward fair assessment in EMI and a cumulative evidence base that can scale across higher education systems.

8. Conclusion

This narrative review synthesized current evidence on fair assessment in EMI and identified practical routines through which content–language separation can be enacted in higher education. The synthesis indicates that fairness is strengthened when disciplinary constructs are stated explicitly, when linguistic demand is treated primarily as an access condition rather than a default grading criterion, and when routine procedures at design, scoring, and moderation keep the intended construct in view; under these conditions, interpretations become more stable and opportunities to demonstrate disciplinary attainment widen for multilingual cohorts. Taken together, the findings position content–language separation as a central organizing principle for fair assessment in higher education EMI.

The review contributes to higher education EMI by linking design, scoring, moderation, and policy to a single organizing principle and by indicating implications for academic staff, programs, institutions, and policymakers. This orientation aligns with Sustainable Development Goals 4 and 10 by decoupling recognition of learning from incidental advantages in English fluency through transparent treatment of content and language. Future research should test these routines through longitudinal studies or assessment-design experiments in higher education EMI contexts. Experimental and longitudinal research can strengthen evidence on how content–language separation affects fairness over time. With these elements in place, content–language separation can move from an abstract ideal to an embedded habit to support assessment practices in EMI that reflect what students know and can do, and thereby consolidate fairer outcomes across higher education.

9. References

Ait-Hroch, A., El Gazi, S., & Ibrahimi, A. (2025). The constructive alignment of objectives, tasks, and assessment in hybrid language training. *World Journal of Advanced Engineering Technology and Sciences, 15*(03), 2606-2627. https://doi.org/10.30574/wjaets.2025.15.3.1182

Aizawa, I., Rose, H., Thompson, G., & McKinley, J. (2025). Content knowledge attainment in English medium instruction: Does academic English literacy matter? *Language Teaching Research*. https://doi.org/10.1177/13621688241304051

Al Fraidan, A. (2024). Beyond the bubble: Unveiling the multifaceted landscape of test wiseness and their operationalization among English-language majors. *Theory and Practice in Language Studies*, *14*(6), 1735–1744. https://doi.org/10.17507/tpls.1406.14

Ardoin, S. P., Binder, K. S., Novelli, C., & Robertson, P. L. (2024). The common element of test taking: Reading and responding to questions. *School Psychology, 40*(5), 607–613. https://doi.org/10.1037/spq0000671

Asquith, S. (2022). An investigation into the roles of guessing and partial knowledge in the Vocabulary Size Test. *TESL-EJ*, *26*(3). https://doi.org/10.55593/ej.26103a15

Baumeister, R. F., & Leary, M. R. (1997). Writing narrative literature reviews. *Review of General Psychology, 1*(3), 311–320. https://doi.org/10.1037/1089-2680.1.3.311

Bozbıyık, M., Balaman, U., & Işık-Güler, H. (2024). Displays of co-constructed content knowledge using translanguaging in breakout and main sessions of online EMI classrooms. *Linguistics and Education, 80*, Article 101275. https://doi.org/10.1016/j.linged.2024.101275

Cade, A. E., & Meuller, N. (2024). Measuring the quality of the OSCE in a chiropractic programme: A review of metrics and recommendations. *The Journal of Chiropractic Education*, *38*(1), 9–16. https://doi.org/10.7899/jce-22-29

Chambers, L., Vitello, S., & Vidal Rodeiro, C. (2024). Moderation of non-exam assessments: a novel approach using comparative judgement. *Assessment in Education: Principles, Policy & Practice*, *31*(1), 32–55. https://doi.org/10.1080/0969594x.2024.2313237

Chapelle, C. A. (2021). *Argument-based validation in testing and assessment*. SAGE. https://doi.org/10.4135/9781071878811

Choi, J., French, M., & Ollerhead, S. (2020). Introduction to the special issue: Translanguaging as a resource in teaching and learning. *Asian Journal of Applied Linguistics, 3*(1), 1–10. https://doi.org/10.29140/AJAL.V3N1.283

Crain, P., & Bailey, B. P. (2022). Easier said or easier done? Exploring the relative merits of common feedback presentations. *Proceedings of the ACM on Human-Computer Interaction*, *6*(CSCW1), 1–19. https://doi.org/10.1145/3512933

De-la-Peña, C., & Luque-Rojas, M. J. (2021). Levels of reading comprehension in higher education: Systematic review and meta-analysis. *Frontiers in Psychology*, *12*, Article 712901. https://doi.org/10.3389/fpsyg.2021.712901

Dimova, S., & Kling, J. (2022). Emerging assessment needs and solutions in EMI in higher education. *Journal of English-Medium Instruction*, *1*(2), 137–152. https://doi.org/10.1075/jemi.00002.edi

Gonsalves, C. (2023). Knowledge of language in rubric design: A systemic functional linguistics perspective. In C. Gonsalves & J. Pearson (Eds.), *Improving learning through assessment rubrics: Student awareness of what and how they learn* (pp. 190−211). IGI Global Scientific Publishing. https://doi.org/10.4018/978-1-6684-6086-3.ch011

Graham, K. M., & Eslami, Z. R. (2020). Does the simple view of writing explain L2 writing development? A meta-analysis. *Reading Psychology*, *41*(5), 485–511. https://doi.org/10.1080/02702711.2020.1768989

Green, B. N., Johnson, C. D., & Adams, A. (2006). Writing narrative literature reviews for peer-reviewed journals: Secrets of the trade. *Journal of Chiropractic Medicine*, *5*(3), 101–117. https://doi.org/10.1016/S0899-3467(07)60142-6

Gronchi, M. (2024). Language assessment in EMI: unravelling the implicit–explicit dichotomy. *Educational Linguistics*, *3*(2), 238–257. https://doi.org/10.1515/eduling-2023-0011

Gülle, T., & Bayyurt, Y. (2024). Translanguaging in content assessment: Voices, experiences and practices of EMI university students. *Language Learning in Higher Education*, *14*(2), 313–336. https://doi.org/10.1515/cercles-2024-0021

Hidayati, N. O., & Andriyanti, E. (2025). Validating a diagnostic reading test for junior high school EFL learners in Indonesia's English massive program using QUEST. *Al-Lisan: Jurnal Bahasa*, *10*(2), 270–283. https://doi.org/10.30603/al.v10i2.6708

Holzknecht, F., Guggenbichler, E., Zehentner, M., Yoder, M., Konrad, E., & Kremmel, B. (2022). Comparing EMI university reading materials with students' reading proficiency: Implications for admission testing. *Journal of English-Medium Instruction, 1*(2), 180–203. https://doi.org/10.1075/jemi.21006.hol

Hou, Z., Zhang, J., Jadallah, M., Enriquez-Andrade, A., Tran, H. T., & Ahmmed, R. (2024). Translanguaging practices in global K–12 science education settings: A systematic literature review. *Journal of Research in Science Teaching, 62*(1), 270–306. https://doi.org/10.1002/tea.22008

Inbar-Lourie, O. (2022). EMI programs and formative assessment. *Journal of English-Medium Instruction*, *1*(2), 204–231. https://doi.org/10.1075/jemi.21014.inb

Iskandarova, G. (2024). Current issues in language assessment and language assessment research and its implication. *Baltic Journal of Legal and Social Sciences, 3*, 228–231. https://doi.org/10.30525/2592-8813-2024-3-24

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Karanfil, T., & Neufeld, S. (2020). The role of order and sequence of options in multiple-choice questions for high-stakes tests of English language proficiency. *International Journal of Applied Linguistics and English Literature, 9*(6), 110–129. https://doi.org/10.7575/aiac.ijalel.v.9n.6p.110

Karnas-Haines, C. (2021). Making assessment actionable through assessor training: A tool for building trust through moderation and calibration. *Grand Challenges in Assessment, 2*(3). https://doi.org/10.61669/001c.24570

Li, H., Zhang, S., & Tang, X. (2024). Effects of test-taking strategy and lexico-grammatical ability on L2 local-level reading comprehension. *Reading in a Foreign Language*, *36*(1), 1–37. https://doi.org/10.64152/10125/67474

Lin, A. M. Y. (2023). Can the Monkey King break through the 'Jin-Gang-Quan' (金剛 圈)? Overcoming the multiple contradictions in EMI education. *Language and Education, 38*(1), 139–147. https://doi.org/10.1080/09500782.2023.2284802

Lo, Y. Y., & Leung, C. (2022). Conceptualising assessment literacy of teachers in Content and Language Integrated Learning programmes. *International Journal of Bilingual Education and Bilingualism*, *25*(10), 3816–3834. https://doi.org/10.1080/13670050.2022.2085028

Lumban Raja, V. (2020). Test item analysis of reading comprehension examination faculty of teachers and training education. *Kairos English Language Teaching Journal, 4*(1), 52–65. https://doi.org/10.54367/kairos.v4i1.847

Macaro, E., Curle, S., Pun, J., An, J., & Dearden, J. (2018). A systematic review of English medium instruction in higher education. *Language Teaching, 51*(1), 36–76. https://doi.org/10.1017/S0261444817000350

Malmström, H., Pecorari, D., & Shaw, P. (2025). Development of academic vocabulary knowledge during English-medium instruction. *Ibérica*, *49*, 157–180. https://doi.org/10.17398/2340-2784.49.157

Mayahi, N., & Jalilifar, A. R. (2022). Self-denigration in doctoral defense sessions: Scale development and validation. *ESP Today, 10*(1), 43–70. https://doi.org/10.18485/esptoday.2022.10.1.3

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13−104). American Council on Education/Collier Macmillan. https://archive.org/details/educationalmeasu0000unse_3ed

Middleton, R., Lewer, K., Antoniou, C., Pratt, H., Bowdler, S., Jans, C., & Rolls, K. (2024). Understanding the processes, practices and influences of calibration on feedback literacy in higher education marking: A qualitative study. *Nurse Education Today, 135*, Article 106106. https://doi.org/10.1016/j.nedt.2024.106106

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). Design and analysis in task-based language assessment. *Language Testing*, *19*(4), 477–496. https://doi.org/10.1191/0265532202LT241OA

Morton, J. K., Northcote, M., Kilgour, P., & Jackson, W. A. (2021). Sharing the construction of assessment rubrics with students: A model for collaborative rubric

construction. *Journal of University Teaching & Learning Practice, 18*(4), Article 9. https://doi.org/10.53761/1.18.4.9

Morton, T. (2022). Using cognitive discourse functions and comparative judgement to build teachers' knowledge of content and language integration for assessment in a bilingual education program. *Journal of Immersion and Content-Based Language Education, 10*(2), 302–322. https://doi.org/10.1075/jicb.21017.mor

Morton, T., & Nashaat-Sobhy, N. (2023). Exploring bases of achievement in content and language integrated assessment in a bilingual education program. *TESOL Quarterly, 58*(1), 5–31. https://doi.org/10.1002/tesq.3207

Motavas, S., & Mahmood, F. (2025, July). *Improving educational equity and outcomes in a first-year engineering programming course through a content and language integrated approach* [Conference presentation]. FYEE 2025 Conference Proceedings, Article 55247. https://doi.org/10.18260/1-2--55247

Murray, N. L. (2022). A model to support the equitable development of academic literacy in institutions of higher education. *Journal of Further and Higher Education*, *46*,1054–1065. https://doi.org/10.1080/0309877X.2022.2044019

O'Donovan, B., Sadler, I., & Reimann, N. (2024). Social moderation and calibration versus codification: a way forward for academic standards in higher education? *Studies in Higher Education*, *49*(12), 2693–2706. https://doi.org/10.1080/03075079.2024.2321504

Ojochegbe, A. T. (2024). Rethinking standardized testing in English language proficiency: Moving toward culturally responsive assessment models. *Jurnal Pendidikan Indonesia*, *5*(12), 1990–1996. https://doi.org/10.59141/japendi.v5i12.6584

Owen, N., & Senel, A. (2025). Enhancing transparency in high-stakes English language assessment: A mixed-methods synthesis of empirical evidence and stakeholder perspectives. *Review of Education*, *13*(2). https://doi.org/10.1002/rev3.70096

Park, J., & Wright, E. A. (2023). Distractor analysis to improve the quality of multiple-choice item development. *English Language Assessment, 18*(2), 73–94. https://doi.org/10.37244/ela.2023.18.2.73

Pearce, J., & Chiavaroli, N. (2020). Prompting candidates in oral assessment contexts: A taxonomy and guiding principles. *Journal of Medical Education and Curricular Development, 7*, Article 2382120520948881. https://doi.org/10.1177/2382120520948881

Pearson, W. S. (2021). Policies on minimum English language requirements in UK higher education, 1989–2021. *Journal of Further and Higher Education*, *45*(9), 1240–1252. https://doi.org/10.1080/0309877X.2021.1945556

Şahan, K., & Şahan, Ö. (2022). Content and language in EMI assessment practices: Challenges and beliefs at an engineering faculty in Turkey. In Y. Kirkgöz, & A. Karakaş (Eds.), *English as the medium of instruction in Turkish higher educatio*n (Vol. 40). Springer. https://doi.org/10.1007/978-3-030-88597-7_8

Sato, T. (2023). Assessing the content quality of essays in content and language integrated learning: Exploring the construct from subject specialists' perspectives. *Language Testing*, *41*(2), 316–337. https://doi.org/10.1177/02655322231190058

Stoeckel, T., McLean, S., & Nation, P. (2021). Limitations of size and levels tests of written receptive vocabulary knowledge. *Studies in Second Language Acquisition*, *43*(1), 181–203. https://doi.org/10.1017/S027226312000025X

Syukur, B. A., & Nurlaily, A. F. (2025). Rasch model-based evaluation of toefl listening items: analyzing difficulty, discrimination, and fit. *Jurnal Smart*, *11*(2), 176–191. https://doi.org/10.52657/js.v11i2.2931

Wang, R. (2021). New perspectives on translanguaging and education [Book review]. *International Journal of Bilingual Education and Bilingualism*, *24*(2), 305–307. https://doi.org/10.1080/13670050.2018.1454043

Watari, T., Koyama, S., Kato, Y., Paku, Y., Kanada, Y., & Sakurai, H. (2022). Effect of moderation on rubric criteria for inter-rater reliability in an objective structured clinical examination with real patients. *Fujita Medical Journal*, *8*(3), 83–87. https://doi.org/10.20407/fmj.2021-010

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Palgrave Macmillan. https://doi.org/10.1057/9780230514577

White, M., & Ronfeldt, M. (2024). Monitoring rater quality in observational systems: Issues due to unreliable estimates of rater quality. *Educational Assessment, 29*(2), 124–146. https://doi.org/10.1080/10627197.2024.2354311

Wudthayagorn, J. (2025). Revisiting English-in-Education policies in Thailand: Ambitious goals, contradictory outcomes. *LEARN Journal: Language Education and Acquisition Research Network*, *18*(2), 1–13. https://doi.org/10.70730/wcsz5574

Xin, K., & Yap, T. T. (2025). Balancing language learning with translanguaging: Insights from Yunnan Agricultural University. *Jurnal Arbitrer*, *12*(1), 27–39. https://doi.org/10.25077/ar.12.1.27-39.2025

Yousefpoori-Naeim, M., Bulut, O., & Tan, B. (2023). Predicting reading comprehension performance based on student characteristics and item properties. *Studies in Educational Evaluation, 79*, Article 101309. https://doi.org/10.1016/j.stueduc.2023.101309