Language Teaching (2025), 1–28
doi:10.1017/S0261444825100876

STATE-OF-THE-ART REVIEW

Fairness, justice, and criticality: Reviewing second language writing assessment

Lia Plakans[1] and Kwangmin Lee[2]

[1]University of Iowa, Iowa, IA, USA and [2]Western Michigan University, Kalamazoo, MI, USA
Corresponding author: Lia Plakans; Email: lia-plakans@uiowa.edu

(Received: 31 July 2025; accepted: 2 August 2025)

Abstract

This State-of-the-Art review examines second language (L2) writing assessment research over the past 25 years through a framework of fairness, justice, and criticality. Recognizing the socio-political implications of assessment, the authors argue for a shift toward more equitable and socially conscious approaches. Drawing from a corpus of 869 peer-reviewed articles across leading journals, the review identifies five major themes: (1) features of writing performance, (2) rating and scoring, (3) integrated assessment, (4) teacher and learner perspectives, and (5) feedback. Each theme is reviewed for foundational findings, then critiqued through questions related to fairness and justice using a critical lens. The authors advocate for a multilingual turn in writing assessment, greater attention to teacher and student voices, and questioning dominant norms embedded in assessment practices. The review concludes with a call for future research to engage with fairness, justice, and criticality in both theory and practice, ensuring that writing assessments serve as tools for empowerment rather than exclusion.

Keywords: fairness; justice; research; writing assessment

1. Introduction

Over the past several decades, there has been a growing critique calling for the need for a more critical lens to acknowledge the power dimensions in language testing (Lynch, 2001; McNamara & Knoch, 2019; McNamara & Roever, 2006; McNamara & Ryan, 2011; Roever & Wigglesworth, 2019; Schissel & Khan, 2021; Shohamy, 1998, 2001). Scholars cite the responsibility to routinely establish fairness, to center justice, and uncover injustice stemming from consequences of assessments (Randall, 2021). This centering requires acknowledging the socio-political contexts in which assessment exists (Schissel & Khan, 2021). For instance, language assessments are used in making decisions about employment, citizenship, education access, and salary, all of which decidedly impact the well-being of people and have material consequences for society. This creates an ethical imperative for our field.

Addressing the issue of equity in assessment research is critical as research informs the applied work of test development and use. Inadequate discussion of equity can lead to harmful effects including limited access and opportunities for language learners. Specializations within the field, such as writing assessment, require critical interrogation as well (Inoue, 2015). With the premise that research should inform our practice, the focus of this State-of-the-Art review is to survey research in writing assessment and analyze it with a framework of fairness, justice, and criticality. The concerns of

https://doi.org/10.1017/S0261444825100876 Published online by Cambridge University Press

fairness and justice can be viewed through criticality, a guiding critical lens, to study and illuminate equity/inequity. Our review will narrow to focus on research that centers writing assessment in teaching and learning in design, analysis, findings, or implications (i.e. not reviewing writing assessment in other contexts such as employment or citizenship). This approach views assessment as part of pedagogy and a tool in developing writing.

In this article, we present themes identified in reviewing a sample of writing assessment research in teaching and learning over the past 25 years and argue for layering a framework of fairness, justice, and criticality over future work in this area to consider how equity has been researched and could be further interrogated. The review speaks to a need to update past reviews of writing assessment research (Hamp-Lyons, 2002; Yancey, 1999). Reviewing knowledge gained from these existing themes in scholarship serves as a starting point to consider where there is further work needed to connect these themes to issues of equity. We propose that writing assessment research faces issues of fairness and justice more intentionally.

2. A framework for fairness, justice, and criticality in language assessment

In this review, we consider a framework that comprises fairness, justice, and criticality as part of a socially conscious approach to writing assessment research. We propose that innovation in writing assessment – through new constructs, methods, and perspectives – will support the field in addressing concerns of fairness and justice. Figure 1 outlines the critical framework applied in this State-of-the-Art review. Assessment context is the foundation on which the interlaced concepts of justice and fairness reside. The framework includes validity, while not the primary focus of our critique, as a dominant quality feature in assessment and a core element that intersects with fairness and justice. Validity includes consequences of an assessment (Messick, 1989) and is essential in work to examine assumptions or dominant narratives (Randall, 2021). We propose a framework that asks critical questions (a critical lens) about fairness and justice that intersect with validity. We will briefly define each of these concepts in relation to writing assessment.

Fairness can be defined as the somewhat procedural considerations in language assessment that impact the test's quality (i.e. construct relevance) and the experience of test takers (McNamara & Ryan, 2011). Fairness is discussed with high stakes standardized assessments, but it is equally important in assessments used in teaching and learning. The linkage between validity and fairness is not new (Messick, 1989), and the understanding of fairness in language testing has evolved greatly, becoming more nuanced (Kunnan, 2014; McNamara & Knoch, 2019; Poe & Elliot, 2019). Of the three terms

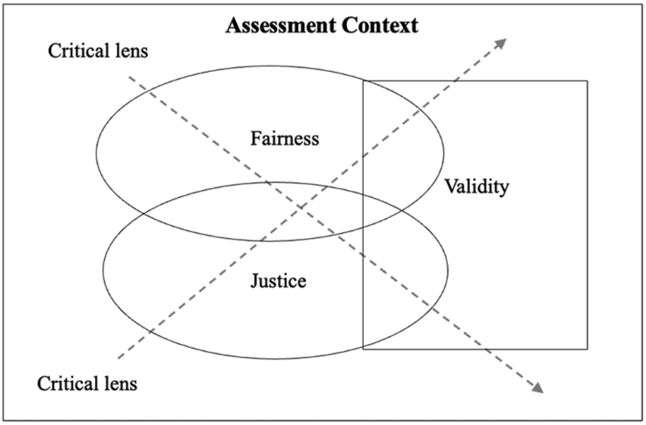

Figure 1. Framework for critical review.

https://doi.org/10.1017/S0261444825100876 Published online by Cambridge University Press

(fairness, justice, criticality), fairness has received the most attention in language assessment but our understanding is also decidedly incomplete. While some have held a broader view of fairness, the definition we are referencing situates fairness closely in the design and delivery of an assessment (McNamara & Ryan, 2011).

Poe and Elliot (2019) reviewed 73 articles on assessing writing to investigate how they included fairness. The researchers identified trends from these articles, namely discussions about minimizing bias, establishing validity, recognizing social contexts, providing legal definitions, and determining ethical responsibility. Their analysis of the representation of these trends over time revealed differing degrees of attention. The technical aspects of fairness – bias and validity – have been increasingly brought up in publications, while social aspects, context, and ethics are mentioned less often. Despite this respectable number of articles, Poe and Elliot felt the scholarship has not 'resulted in shared taxonomies across disciplinary orientations, led to a deepening of theoretical conceptualizations of fairness or brought about innovative classroom approaches' (p. 15). The field needs to remedy this gap and address social and ethical aspects of fairness more deeply.

Justice can be defined as the socio-political and consequential implications of testing. In some definitions, justice is combined under fairness in relation to consequences (Kunnan, 2014); however, we consider justice to be overlapping with but separate from fairness (McNamara & Ryan, 2011). Justice requires us to examine inequities in the use of assessments and their impact on individuals, communities, and society. Randall (2021) speaks to the need to apply a justice-oriented framework of anti-racism in rethinking constructs in educational measurement. In doing so, we interrogate the purpose of an assessment in terms of justice as well as reflect on our positionality as assessment users/developers/researchers/teachers. A justice-oriented approach seeks the input and understanding of stakeholders in the assessment process.

Validity has been given considerable attention in writing assessment over the past 25 years or more (White, 2019) and can be described as the quality of a test's content coverage and relevance to the construct being assessed. According to Kane, 'Validity is not the property of a test. Rather, it is a property of the proposed interpretations and uses of test scores' (Kane, 2013, p. 3). In education, this perspective considers validity in assessments as supporting the assessment-learning cycle rather than solely to judge learning or learners (Shepard, 2016). It also subsumes consequences of tests, which directs writing assessment research to engage with fairness and justice.

Relationships between these key concerns in writing assessment have been discussed in the field. While different views have appeared, as stated previously, we are following the work of McNamara and Ryan (2011), who have drawn upon Messick (1989) and point to his 'Facets of Validity' as key to fairness, citing construct validity and test use (relevant utility), while value implications and social consequences align with justice. McNamara and Ryan (2011) conceive of justice as an umbrella encompassing fairness, values, and consequences. That said, the relationship with fairness, and arguably, with validity, is reciprocal: 'Concerns for fairness [...] have the paradoxical potential to cloud issues of justice' (McNamara & Ryan, 2011, p. 175). In assessment, validity has been a major linchpin for research and practice; however, fairness has recently risen as an equal partner to validity. Fairness is a necessary criterion for validity. For example, when assessments advantage certain groups over others, this constitutes construct irrelevance – a concern of validity discussed by Randall (2021) – but also a threat to fairness.

Justice has received limited attention to the point that Schissel and Khan (2021) have challenged the field:

> The current state of (limited) engagement with social issues does not make the issues disappear nor diminish their significance, but rather simply reflects an ignorance, and perhaps even cowardice, and works to limit, stifle, and police a field at precisely the time when it could be developing in other ways. It is time to work actively against exclusionary disciplinary approaches (p. 2).

https://doi.org/10.1017/S0261444825100876 Published online by Cambridge University Press

To interrogate issues of power, **assessment context**, as the socio-political milieu, becomes a necessary consideration. Social/socio-political factors might also include dominance of the language of an assessment; cultural knowledge or practices embedded in an assessment that are unfamiliar to test takers; test purposes such as gatekeeping; or assessment tied to advancement and opportunity. For example, assessments used in citizenship or employment requirements will garner somewhat different critical questions than assessments in educational settings. Context also determines the stakeholders impacted by an assessment and the decisions made therein. In this review, our focus is on writing assessments used in education, narrowing the critique to research related to teaching and learning writing. While the many domains in which writing assessments are used deserve attention, our experience has been primarily as language educators, and thus this is the focus we feel best qualified to critique. Further, this focus is most relevant to *Language Teaching*.

Criticality entreats us to ask questions about fairness and justice and about context and power. It provides a process through a critical lens to challenge us in enacting a justice-oriented approach. Criticality leads scholars to question established and invisible practices and systems that perpetuate inequity in our society, communities, and classrooms (De Costa, 2018; Kubota & Miller, 2017; Soto, 2022). It sheds light on power, dominance, and oppression. Criticality leads researchers to ask questions such as, 'Whose standards are these?', 'Who is excluded?', and 'Does this uphold hegemony and marginalization?' To do this work, researchers need to consider their own positionality and reflect on their experiences, underlying assumptions, and understanding of racism, oppression, inequity, and dominant norms that underlie their work, beliefs, and assumptions (Brown, 2013; Sealey-Ruiz, 2021). Criticality is not an answer but a process that changes our thinking and our vision. While taking a critical view can create dissonance and discomfort, such 'crises' (Kumashiro, 2000) are how we learn, unlearn, and do better. Therefore, it should have a central role in research. We need to use a critical lens in research to understand and to take responsibility for inequity in writing assessment in teaching and learning.

In our review, we are proposing fairness and justice be informed and transformed by criticality in writing assessment. In particular, we advocate for more work to consider the social/socio-political context of writing assessment in teaching and learning. Adopting a perspective of criticality interrogates assumptions and contexts for power and equity (or inequity). This call to language testing is not new. The need for this work resides in the ubiquity of assessments in decision-making and the impact of those decisions. Critical language testing has had a voice in the field since the 1990s, through work by Shohamy (1998, 2001), Lynch (2001), McNamara & Roever (2006), and others. In the words of Shohamy, 'Critical language testing assumes that the act of testing is not neutral. Rather, it is both a product and an agent of cultural, social, political, educational, and ideological agendas that shape the lives of individual participants, teachers, and learners' (1998, p. 332).

3. Writing assessment research in teaching and learning

To review writing assessment critically, context is needed. Asking critical questions requires understanding the purpose and uses of assessment and who is impacted. Thus, for this review, we have narrowed the focus to writing assessment research related to education, specifically, teaching and learning. While this still casts a wide net, centering our review on assessment as part of pedagogy means less attention to assessments used in citizenship requirements, admissions processes, or employment/salary decisions. At one time, language testing research and theory-building focused largely on large scale assessments, due to the high-stakes decisions made using these tests such as employment, education access, and citizenship (Inbar-Lourie, 2008). However, assessment's role in everyday learning in classrooms is equally important in facilitating access and opportunities to learn. Our focus is on inquiry into assessment for instructional purposes, that is, practices that are used to support learning enacted by teachers in educational contexts. We reviewed research on classroom-based assessment and also drew on studies with larger scale assessment that included findings and implications useful to the teaching and learning of writing.

https://doi.org/10.1017/S0261444825100876 Published online by Cambridge University Press

Discussions of fairness, consequences, and ethics have involved large-scale standardized testing due to the high stakes impact and policy dimensions of these assessments. However, assessments in classrooms should also be considered critically as they can affect the access, trajectory, and opportunities of learners. Potentially, since these assessments are subject to less scrutiny and generally do not adhere to established quality checks around issues of equity, they may cause greater harm. Asking critical questions in research should further the expectations for fairness and justice for learners in these contexts.

Scholars in the field of writing assessment have been increasingly interested in what kinds of assessments promote learning, which methods are effective and efficient in classroom assessment, and how to productively prepare teachers to use assessments (Crusan et al., 2016). Most recently, the conceptualization of the student in the center of assessment in learning has been detailed by the learner-oriented assessment (Gebril & Brown, 2019; Jones & Saville, 2016; Turner & Purpura, 2016). In a review of 25 years of research in writing assessment, White (2019) described the shifts in thinking about writing assessment and classroom applications: 'While different community positions remain distinct – writing teachers are unlikely to love multiple-choice tests and test professionals will remain distrustful of formative teacher evaluations – the hardened positions of yesterday have become more flexible' (p. 42).

In this State-of-the-Art article, we present a review of published research to understand themes in assessing the skill of writing. We explore its interface with teaching and learning and put forward areas for more work by applying a framework for critique related to fairness, justice, and criticality (see Fig. 1). This review leads to the discussion of potential future innovation and challenges for writing assessment in language education.

4. Methodology

Our approach to this review was multifaceted, aiming to provide both breadth and depth. The steps in our synthesis of the research are briefly listed below:

- Collect and vet research articles across specialized high-impact journals in L2 writing, assessment, and language education
- Conduct initial review of articles for patterns and themes
- Evaluate connections to teaching and learning to narrow the pool
- Explore emergent topics within these areas of writing assessment
- Consider implications from this review for pedagogy
- Apply framework (Fig. 1) for critical review of themes for future directions.

Our search began with a range of journal issues published between 2009 and 2025. We entered the keywords 'writing' or 'writing assessment' in multiple databases to extract and form a unified body of relevant studies. This initial search yielded a great many articles, beyond the scope of feasibility for the article; thus, we decided to narrow the journal inclusion base. The review focused on major flagship journals in research on language assessment, writing, and second language learning. The journals range in impact factor (JIF) from 1.4 to 5.0 and are listed in the Social Sciences Citation Index (SSCI, Web of Science). These criteria were used to assure that the research reviewed has undergone a competitive peer review process. The journals from which studies were sourced include *Language Teaching, TESOL Quarterly, Journal of Second Language Writing, Assessing Writing, Language Assessment Quarterly, Language Testing, Journal of English for Academic Purposes, System*, and *Journal of English for Specific Purposes*. However, we recognize that there are many other journals, books, and book chapters that include important and diverse research perspectives. For this study, we kept the scope narrow to allow for depth but feel that further study that casts a wider net would be fruitful. These journals were searched for articles using the keywords 'writing + assessment', which

Table 1. Articles reviewed on writing assessment in teaching and learning research

Journal	Count
Assessing Writing	456
System	109
Journal of Second Language Writing	71
TESOL Quarterly	51
Journal of English for Specific Purposes	44
TESOL Journal	39
Language Testing	39
Language Assessment Quarterly	25
Language Teaching and Research	30
Journal of English for Academic Purposes	5
Total articles reviewed	**869**

led to an initial sample of over 800 articles. Table 1 presents the count of articles reviewed in each of the seven journals.

With these targeted journals in the areas of language teaching, language learning, writing, and assessment, we conducted a comparative process to review for emergent themes by grouping similar journals. This comparative approach allowed us to consider broad themes in several areas. In the first round, we searched in *Language Teaching, System*, and *TESOL Quarterly*, journals known to publish on a wide range of topics related to language teaching and learning. For the next stage of the search, we considered journals that published articles focusing primarily on L2 writing or L2 writing assessment: *Journal of Second Language Writing, Journal of English for Academic Purposes, Journal of English for Specific Purposes*, and *Assessing Writing*. Lastly, we shifted our attention to journals that publish articles exclusively in language assessment: *Language Assessment Quarterly* and *Language Testing*. Following this review, we sampled two journals in educational assessment and measurement, *Educational Assessment* and *Assessment in Education*, but neither published any articles that focused on second language writing.

Through this step-by-step process, we reviewed research studies to generate an initial list of themes in writing assessment research. These themes were rather rough, with both overlap and breadth. The following was the list of themes: (1) writing assessment in general, (2) rating and scoring, (3) features of writing performance, (4) process and strategies, (5) integrated assessment, (6) assessment, teaching, and learning, (7) teachers' perspectives, and (8) learners' perspectives. To make the review and critique more meaningful, we narrowed the pool of research articles and themes further to address context. Thus, we scored all the articles in terms of relatedness to teaching and learning.

To ensure we were selecting the most relevant research for the context of interest, we developed a process to rate the studies by their relevance to teaching and learning. Each reviewer read a set of 20 articles independently using a simple three-point scale to score relevance. After this initial rating, we compared scores to discuss the differences between levels (i.e. what was a 1 versus a 2 or 3) as concretely as possible, which led to further refinement. We also assigned an anchor example for each score level for our reference and calibration. Then, each reviewer gave each article a score on the scale (shown below) to reflect how closely the content connected to teaching and learning either in its research design, analysis, or implications.[1]

> Each article will be given a score on a three-point scale to reflect **how closely it connects L2 teaching and learning with assessment,** either in its research design, analysis, or implications of the findings.

https://doi.org/10.1017/S0261444825100876 Published online by Cambridge University Press

(1) not related or only mentioned in implications or rationale
Example: Ling (2017). Are TOEFL iBT® writing test scores related to keyboard type? A survey of keyboard-related practices at testing centers. *Assessing Writing, 31*, 1–12.

(2) somewhat related, appearing in at least one area (design, analysis, or discussion)
Example: Huang (2012). Using generalizability theory to examine the accuracy and validity of large-scale ESL writing assessment. *Assessing Writing, 17*(3), 123–139.

(3) assessment is highly connected with teaching and learning throughout the article's framing and in the research design and analysis
Example: Vincelette and Bostic (2013). Show and tell: Student and instructor perceptions of screencast assessment. *Assessing Writing, 18*(4), 257–277.

From this rating, we selected research articles that were strongly connected (scored 3) or somewhat connected (scored 2) to teaching and learning, and we excluded articles less connected to these areas of interest (scored 1). In the reference section, articles included in the final review are marked with asterisks (*).

We read and analyzed this narrowed pool of 69 articles, identifying emergent themes in the studies and reviewing their findings with a critical lens. Coding strategies from qualitative research (Saldaña, 2011) were used to create initial and axial codes. We conducted the initial coding individually, before discussing codes, patterns, and potential themes together. Some themes stayed the same as the initial round of comparative coding. Others were combined into different themes. Frequently, an article fell across themes, and we discussed how to distinguish themes and articulate the core concepts of articles.

In our review of research since 2009, five major themes emerged in published writing assessment research: (1) features of writing performance; (2) rating and scoring; (3) integrated writing assessment; (4) teachers' and learners' knowledge, beliefs, and perspectives; and (5) corrective feedback. These themes emerged as central overarching areas of research attention during this time period. Sub-themes within each of these major themes emerged and will be presented in the next section.

These five areas have generated a substantive amount of attention n answering empirical and interpretive questions, resulting in foundational knowledge. A need for more attention to fairness and justice became apparent, however, requiring critical approaches as a direction to elevate this work. A final stage of analysis led us to read the review of research in each theme with a critical lens to consider where authors addressed fairness, justice, or criticality, as drawn from our framework (Fig. 1). This critical review of the research was a second layer of analysis to considering the major themes through our framework and implications for future research.

Validity and context are important parts of the framework (Fig. 1); however, they do not emerge as central findings in the critique for different reasons. Writing assessment research has a robust tradition of research questions related to validity, and thus it appears throughout the themes described in the review. However, interest has not focused on how validation has or could grapple with the challenges of fairness and justice in writing assessment using a critical lens. Context, in contrast to validity, has not been a central feature of writing assessment research and thus does not emerge as strongly in the identified themes, yet it impacts fairness and justice in important ways. We recognize a degree of interpretation in this process and acknowledge this organizational scheme is filtered through our own perspectives of writing assessment and research.

5. Themes in writing assessment in teaching and learning

In this section, a synthesis of research organized by themes is reviewed to highlight recent research in writing assessment related to teaching and learning. Using our framework for critical review, each

https://doi.org/10.1017/S0261444825100876 Published online by Cambridge University Press

theme is followed by a discussion of (1) whether or how the research addresses fairness and justice in assessment and (2) where there is space or potential for a critical lens in this work.

5.1. Features of writing performance

Researchers in L2 writing assessment have made major efforts to understand how specific textual characteristics in L2 writing relate to test scores and second language proficiency. Some of the most investigated language features include complexity, accuracy, lexis, and fluency (CALF) (Adams et al., 2015; Plakans et al., 2019; Wolfe-Quintero et al., 1998). However, L2 writing assessment scholarship has also expanded to incorporate other measures to investigate the quality of writing, rather than relying exclusively on CALF. For example, Taguchi et al. (2013) examined college-level ESL learners' argumentative essays, focusing in particular on LANGUAGE USE and CONTENT. Language use was operationalized as 'facility in the use of effective, complex constructions, and few or no grammatical errors' (p. 423), and CONTENT referred narrowly to the accurate representation and effective use of source text. While both were found to be indicative of higher essay scores, the content measure contributed more to language use than total score variance. The finding that content accounted for more variance in total score than language use suggests that content may not be merely a surface-level construct. Instead, it is possible that content serves as a higher-order construct that shapes how language is used. More recently, Sato (2024) clarified the construct of content as crucial yet absent information in the context of content and language integrated learning, highlighting the inextricable nature of language and content.

Another framework for writing performance features was proposed by Kuiken and Vedder (2017), who highlighted the importance of assessing functional adequacy in L2 writing. Informed by Grice (1975)'s maxims of quantity, relevance, manner, and quality, the researchers developed a rating scale scoring the dimensions of CONTENT, TASK REQUIREMENTS, COMPREHENSIBILITY, COHERENCE, and COHESION. Based on essays written by L2 Dutch and Italian learners, the researchers subjected the scale to correlations and reliability checks. Intra-class correlations are estimated to range from acceptable to excellent among four different dimensions. Furthermore, the raters' judgments of the writings composed by the same participants were consistently scored, with reliability estimates calculated between .455 (task requirements for Italian) and .877 (comprehensibility for Dutch). This study contributed to an understanding that L2 writing is a multidimensional construct that requires the orchestration of not only linguistic but also pragmatic resources to communicate and make meaning possible.

Recent research applying CALF measures aims to examine how such measures vary as a function of differing conditions of writing. For example, seeking to examine the relationship between task complexity and CALF measures, Frear and Bitchener (2015) found that an increase in reasoning demands and the number of task complexity elements was facilitative in eliciting better lexical complexity and syntactic complexity. Adams et al. (2015) analyzed the performance of 96 students on engineering simulation tasks through text chat. They manipulated two different task conditions: the amount of direction to complete a task and the language support required. The findings suggested that more complex tasks push learners to produce more accurate language. A study by Ong and Zhang (2010) examined the effects of providing planning time, ideas and macro-structure, and draft availability. Learners given planning time, ideas, and a macro-structure were found to produce more complex language with more words (i.e. fluency). In the context of multi-skill assessment involving writing based on reading and listening passages, Plakans et al. (2019) found that fluency, operationalized as total word count, was the strongest predictor of integrated writing proficiency. Morphological accuracy contributed more to score variance than either syntactic or grammatical accuracy. Although significant, the contribution of complexity, operationalized as mean length of T-unit to score variance, was not as strong as fluency and accuracy. To summarize, CALF is informative but not a fixed linguistic system of learners/test taker's language, as research has shown; it is subject to variation due to a host of reasons, such as those reviewed here.

https://doi.org/10.1017/S0261444825100876 Published online by Cambridge University Press

While research on the measurement of writing features in L2 writing performances over the past decade has mostly been carried out using the CALF framework, a growing number of research studies draw on other features that speak to capabilities in higher-order thinking, including authorial voice, critical thinking, and source use (Behizadeh & Engelhard, 2014; C. G. Zhao, 2013; Kim & Crossley, 2018). L2 writing appears to be a multidimensional language skill that subsumes and interacts with various types of capabilities beyond lexicogrammatical features.

Over the past decade, research on writing features in assessment has not specifically centered on classrooms; however, the composite of findings holds implications for teaching and learning. The research suggests that instruction should address the multidimensional nature of second language writing. Writing assessment used in teaching should provide students and instructors with insights, not just on the accuracy, fluency, or complexity of students' writing, but also into how students engage with content, they meet the demands of the task, and features such as coherence and cohesion that are used to increase the comprehensibility of their writing. Another implication is how the design of assessment tasks impacts the features in the writing performance of students. For example, careful consideration is needed for the complexity of tasks as well as what supports are provided, including planning time and source material.

5.1.1. Fairness, justice, and criticality in writing features research

The current approach in researching characteristics of writing at different proficiency levels needs a broad critical lens to uncover embedded norms and values. From a validity standpoint, we need to interrogate the constructs and standards that have highlighted these features for attention. Shohamy stated in 'Critical language testing and beyond' that we need to specify 'whose knowledge the tests are based' on (1998, p. 333). A critical approach requires recognition and transparency on where the characteristics, such as CALF, come from, as well as the interpretation of quality based on specific measures. What defines proficiency and proficiency levels in English or any language on which L2 writing is assessed? On what variety of English are measures of accuracy determined? Why is fluency a predictor when conciseness is also a value of communication?

Questions related to justice need to be asked about power, equity, and fairness in relation to how we characterize writing performances in assessments. These are big questions, each with an individual research agenda. From a multilingual rather than a monolingual perspective, CALF troubles this tripartite classification and could lead to more equitable measures. It is promising that scholars who are beginning to look at sociocultural features of writing are adding depth to the traditional features that have been the focus of L2 writing (Behizadeh & Engelhard, 2014; C. G. Zhao, 2013). These new areas for research are shifting focus from accuracy and length to communicative impact and content (Kim & Crossley, 2018; Kuiken & Vedder, 2017). These alternative targets allow, or could allow, a more just consideration of language varieties in writing assessment. The field perpetuates systems that subjugate multilingual writers if we ignore the issue of dominant norms in 'features' or 'proficiency levels' in our research.

5.2. Rating and scoring L2 writing

As L2 writing assessment has shifted from indirect measurement of writing knowledge to direct measurement of writing performance (Crusan, 2010), challenges of rating have become a source of concern for researchers and practitioners. This work focuses in on rating and scoring of performances in writing assessments. In the context of L2 writing assessment, research has shown that variables such as topic, discourse mode, genre, and time limits can pose a threat to reliable rating, thus interfering with accurately reflecting the student's proficiency and the potential for providing meaningful use of scores (Schoonen, 2005).

The rating scale, a critical tool that informs 'decisions and inferences about writers' (Weigle, 2002, p. 108), plays a central role in L2 writing assessment. Its systematic and rigorous nature aligns with the

https://doi.org/10.1017/S0261444825100876 Published online by Cambridge University Press

broader goals of educational research, which seeks to generate evidence-based findings grounded in transparent evaluative practices. Research reviewed has addressed rating scales primarily with regard to their development, adaptation, and validation. The recognition that writing is a multifaceted construct that encompasses not only linguistic but sociolinguistic and pragmatic aspects of language has led researchers to develop rating scales incorporating complex nuanced constructs. For example, attempts have been made to develop analytic rubrics measuring authorial voice in L2 argumentative writing (C. G. Zhao, 2013), authenticity (Behizadeh & Engelhard, 2014), and critical thinking (Saxton et al., 2012). Some studies have introduced scoring rubrics intended for specific modes of writing, such as a rubric for electronic portfolios of L2 writing (Lallmamode et al., 2016) and for reading-to-write tasks (Chan et al., 2015). Studies that report newly developed scales also discuss their validation procedures (Ramineni, 2013). For instance, findings of Saxton et al. (2012) specifically reference inter- and intra-rater reliability as a measure to ensure consistency, which is particularly hard to achieve for a newly implemented rubric.

A well-established rubric, if put to use in the local context, might not work well in distinguishing proficiency levels specific to the students or curriculum. The need for revising a rating scale in accordance with the demands of context is highlighted in Janssen et al. (2015) and Banerjee et al. (2015). These studies emphasize the importance of following both theoretical and empirical criteria in constructing a rating scale appropriate for specific contexts.

In L2 writing assessment, human raters have long been recognized as potentially impacting reliability due to differential application of the rating scale and individual interpretations. Barkaoui (2010) found raters paid selective attention to only language, rhetoric, and ideas when using a holistic scale, whereas an analytic scale led them to attend to all the listed criteria. However, a later study (Winke & Lim, 2015) reported that raters' attention was not equally distributed across categories using an analytic scale; in fact, it was impacted by the sequence in which the criteria were organized in the rubric.

While the notion that rater training is widely accepted as playing a pivotal role in securing reliable ratings of essays (Weigle, 1998), several research studies carried out over the past decade yielded mixed evidence as to the effectiveness of rater training. Implementing a short-term training program, Attali (2016) compared the psychometric properties pertaining to new and experienced raters in assessing writing, such as reliability derived by G-theory and confirmatory factor analysis. Based on the findings that there were only small differences between the two rater groups in terms of scoring consistency and validity, he suggested that rating is influenced by 'learning that occurs during initial training and abilities that are acquired prior to training' (p. 107), rather than rater experience *per se*. The implications of this research study are powerful, such that rater reliability may be more contingent upon the cognitive framing established during initial training, rather than the gradual accumulation of expertise. A similar finding was reported in a longitudinal study by Lim (2011), where the novice raters who initially displayed fluctuation in rating quality fairly quickly caught up with their experienced counterparts. The effectiveness of individualized feedback to raters, a more specific form of rater training, has been found to be inconclusive in the literature. A Rasch analysis carried out by Knoch (2011) found training ineffective in bringing about change to reduce bias or variability of individual raters.

Several large-scale language tests have embraced computerized scoring to achieve higher reliability while reducing the high costs associated with human rating. Recent research, however, has focused on its limitations. One inherent vulnerability of the rating software is that it is not sensitive to detecting construct-irrelevant strategies, commonly referred to as test-wiseness strategies. In Bejar et al. (2014), examinees who substituted a portion of their essay with less frequent and longer lexical choices were given a higher score by the software *e-rater* than essays without such manipulation. Also, the challenges associated with scoring content or higher-order thinking that goes into the composition of essays remain unresolved in automated scoring (Attali et al., 2013; McCurry, 2010). In addition, automated essay scoring by way of artificial intelligence (AI), machine learning (ML),

https://doi.org/10.1017/S0261444825100876 Published online by Cambridge University Press

and natural language processing (NLP) is gaining traction in local contexts of L2 writing assessment. Hannah et al. (2023) developed an ML-supported automated system that measures young students' writing (Grades 3 to 6) on the features of task fulfillment, organization and coherence, and vocabulary and expression. Their findings indicated (1) that the reliability between human raters and the automated system was higher than that comparing only human raters, and (2) that the agreement of human-AES ratings increased with grade. A study by Sickinger et al. (2025) indicated that comparative judgment, which draws on the benefits of both automated scoring and human rating, is high in reliability for both holistic and analytic rating of young learners' essays. Potter et al. (2025) leveraged NLP to operationalize the lexical, syntactic, cohesive, and rhetorical aspects of academic language as produced in source-based argumentative writing by tenth-grade students. Local cohesion – semantic connection between sentences – was not found to have a significant effect on holistic score of source-based writing, whereas source cohesion – semantic cohesion with the source text without verbatim copying – was estimated to have the largest effect on the holistic quality.

An increasing number of L2 writing assessment studies have employed a statistical method called generalizability theory (G-theory) to provide a nuanced understanding about reliability. A G-theory study carried out by Gebril (2010) found that the scoring reliability of an integrated reading-into-writing task is as high as that of an independent writing task. With this finding, the researcher argued for increased use of integrated performance assessment for academic purposes. Also, in the context of integrated writing assessment, Ohta et al. (2018) examined the comparability of the holistic and analytic scales. They found that the analytic scale yields a higher degree of reliability, with one of the sub-scales, *source use*, identified as the most consistent and reliable. Han (2019) carried out a G-theory analysis to examine the extent to which task topic, rater, and direction of interpretation (e.g. English to Chinese vs Chinese to English) are subject to measurement error in a language interpretation assessment. Bouwer et al. (2015) incorporated genre as a potential source of variance and reported that the generalizability of writing scores is contingent on the genre selected in the assessment. Huang (2012) used G-theory in estimating the reliability of writing scores obtained by second language and first language English-speaking students and found systematic differences in the scores depending on the language profile of the test takers. Noting less precision and reliability for scores assigned to the L2 test takers, Huang (2012) directed attention to potential fairness issues inherent in assessing multilingual test takers. Gtheory has potential as practical guidance for test developers and teachers in the classroom interested in understanding reliability and the impact of raters, task type, test takers, and other aspects of performance-based writing assessments.

Our discussion on rating and scoring has been one of securing validity and reliability. Research in the past decade has provided an increasing number of rubrics that reflect the multidimensionality of L2 writing by incorporating aspects such as voice and critical thinking. Those rubrics have great potential for classroom use. Ecological validity should be of utmost concern to teachers who would like to use a well-known, established rubric. Without consideration of contextual idiosyncrasies and limitations, a rubric, however well established for its sound psychometric properties, might not function well in different contexts. Thus, it may be even more important to recognize that consistency when scoring writing can be impacted by the format of a rubric (analytic vs holistic) and by experience with the rubric and level of training in using it. These findings support practices such as teachers scoring student writing together across courses and ongoing professional development on using rubrics and rating writing. Also, reliability has been addressed in terms of interrater consistency and use of generalizability theory as means to identify the amounts of variance supplied by multiple sources. We suggest that, although using automated scoring might help decrease variance in grading, it still lacks a mechanism to examine higher-order skills, such as critical thinking, content representation, and organization.

https://doi.org/10.1017/S0261444825100876 Published online by Cambridge University Press

5.2.1. Fairness, justice, and criticality in rating and scoring research

In research on rating and scoring, attention to reliability has the potential to support fairness, as it should create equal opportunities to all test takers if everyone is scored in the same way. However, a definition of language proficiency is embedded in our scoring scales to establish levels of writing proficiency. These definitions and 'standards' lead to critical questions, including: Whose language varieties or language development patterns are these levels based on? Do they represent or prioritize one community's language variety over others? Researchers have investigated expanding rubrics to include criteria such as authorial voice (C. G. Zhao, 2013) or critical thinking (Saxton et al., 2012), which may push raters beyond traditional language standards; yet these too are not neutral and are imbued with cultural and linguistic values. Research on rating and scoring has not closely scrutinized the potential for racial and linguistic bias and discrimination underlying scoring or rating scales. Scoring scales have been seen as devoid of human influence, largely taken as static documents, but this assumption reveals an inherent misunderstanding of how scales, standards, and rubrics are designed and enacted. Studies of rubrics-in-context provide a pathway to understand the impact of rubrics and the benefits of being dynamic and responsive (Banerjee et al., 2015; Janssen et al., 2015).

Standards and criteria for evaluation are related to another issue in need of more criticality: the bias of raters and teachers. Shohamy points to this issue in discussing the interpretation that comes with scoring, where critical language testing 'considers the meaning of language test scores, the degree to which they are prescriptive, final and absolute, and the extent to which they are open for discussion and interpretation' (1998, p. 333). Studies have found substantial evidence for rater bias in writing assessment, even with training to minimize its influence. It is a recognized weakness of performance assessments (Barkaoui, 2010; Winke & Lim, 2015) although researchers seek to mitigate this problem (Attali, 2016; Knoch, 2011). A critical lens has not been used in deeply interrogating and understanding the power raters have in affecting test outcomes and the potential inequities created by the bias that raters bring to assessment.

One might assume that with automated scoring of writing performances, which has received increased use and attention over the past 20 years (e.g. Attali et al., 2013; Bejar et al., 2014), biases and discrimination would be minimized. Not so. The programming involved in establishing automated scoring depends on both scoring scale characteristics (standards) as well as human scorers to train and calibrate the system. The same biases held by human raters as well as dominant 'native' norms may be programmed into the automated scoring. Automated scoring may only mechanize inequity issues that reside in scoring scales and human raters.

The bias teachers bring to reading student work is an area that would benefit from a critical approach. Ferris et al. (2011) and Ferris (2014) show potential in looking at teachers' philosophies and their actions. This research has identified teachers who adjust feedback to student needs or who seek to give agency to students. This work is a promising direction for writing assessment research. The impact of differences in teachers' experience and knowledge on scoring reliability is important, but equally important are values and power issues embedded in this knowledge and these processes. These are potential areas for more work.

5.3. Integrated writing assessment

Over the past two decades, L2 writing assessment research has paid increasing attention to integrating skills in assessment. These tasks require test takers to perform two or more modalities in a writing task, such as reading-to-write or reading-listening-writing tasks. Three points are often used to justify increasing popularity of integrated assessment. First, integrated assessment features a high degree of authenticity (Gebril & Plakans, 2014; Huang, 2012; Plakans, 2009) since the integration of language skills approximates the way language is used in real life. For instance, test takers who listen to and read academic passages for writing in integrated tasks face similar demands in academic writing contexts. Second, integrated assessment tasks are believed to have a facilitative impact on the teaching and learning of language in the classroom, hence positive washback, that is, learners and teachers focus on improving language skills with a balanced approach in mind. Lastly, students are less likely to need to draw on background knowledge if given the same reading and/or listening source passages to work on.

Research has explored the relationship between independent (i.e. writing only) and integrated skill writing, showing some similarities as well as differences. A G-theory study carried out by Gebril (2009, 2010) yielded similar reliability indices for both types of writing tasks. In a similar vein, in an L2 writing quality model formulated by Kim and Crossley (2018), the lexical, syntactic, and cohesive features of independent and integrated writing did not demonstrate significant differences in internal factor structure. While the findings of Gebril (2009, 2010) and Kim and Crossley (2018) provided convergent evidence that independent tasks and integrated tasks are similar in some respects, other studies highlight differences between the two tasks in terms of cohesive devices used (Tywoniw & Crossley, 2019), phraseological complexity (Zhang & Ouyang, 2023), and fluency (Michel et al., 2020). These studies provide evidence that important information is gained from writing-only tasks, but also from integrated tasks that include skills and source materials in the task.

Investigation into the processes adopted by test takers has emerged as an important aspect of test validation; a valid test should elicit cognitive processes and test strategies as intended (Xu & Wu, 2012). Integrated assessment has also been researched from the perspectives of test-taking strategies (Cohen, 1998). Yang and Plakans (2012) found that, in an integrated reading-listening-writing task, test takers employ various strategies to integrate multiple modalities. Their findings validated a multifaceted model of integrated writing strategy, consisting of positive contributions from discourse synthesis (Asencion, 2008; Spivey, 1984) and self-regulating strategies (Xie, 2015), as well as negative contributions of test-wiseness strategies. Furthermore, a process-oriented study by Michel et al. (2020) found that test takers demonstrated the highest lexical productivity during the first and last stages of testing time for the integrated task, whereas fluency (e.g. mean active writing time) in the independent task was evenly distributed throughout all stages of writing. Therefore, integrated tasks are distinctly different from writing-only tasks, not only in terms of written products, but also in the underlying composing processes.

Caution with integrated assessment persists despite its perceived advantages. In terms of construct validity, the psychometric structure of integrated writing has not been fully understood (Knoch & Sitajalabhorn, 2013). As noted by Cumming (2013), integrated assessment carries the danger of confounding the measurement of writing with the ability to comprehend the source passage. Indeed, researchers diverge on the role of reading in one's performance of integrated writing (e.g. Asencion, 2008; Trites & McGroarty, 2005; Watanabe, 2001).

Concerns have also been expressed with regard to the appropriate use of sources, an integral aspect of integrated writing, and research has consequently followed (Hyland, 2009; Plakans, 2009). Defining inappropriate source use as a test-irrelevant construct in integrated writing, Yang & Plakans (2012) identified a negative, direct impact of verbatim source use on test scores. In a similar vein, Uludag et al. (2019) reported that appropriate source use, operationalized as number of ideas incorporated from the source text and accurate content representation, had positive correlations with integrated writing performance. Furthermore, analyzing integrated writing performance involving both reading and listening, Plakans and Gebril (2013) reported that higher-rated essays were characterized by proper use of the listening passage with the inclusion of important ideas, whereas relatively lower-rated essays displayed heavy reliance on the reading passage, copying words and phrases from it. This finding has been echoed in Kyle (2020) who employed Natural Language Processing to computerize annotation of verbatim source use originating from the reading passage and listening passage. However, Weigle & Parker (2012) argued that extensive borrowing from the source text is not a cause for concern, based on their findings that less than ten percent of the study participants exhibited verbatim use of the source texts. Lee et al. (2025) carried out a series of generalized linear models, estimating a negative relationship between verbatim source use and organizational features. Based on this finding, they

https://doi.org/10.1017/S0261444825100876 Published online by Cambridge University Press

suggested that instructional focus on organization skills (e.g. authorial voice, development of ideas, coherence, organization) might help reduce reliance on verbatim source use.

This area of research holds useful applications for writing assessment in teaching and learning although the studies have not been centered in classroom contexts. For teaching academic writing, which generally requires other language skills (i.e. reading and listening), teaching integration of source material in writing holds potential. Using integrated assessment can provide feedback to students on their strengths and areas for improvement to support this development as academic writers. Research shows that these tasks have similar capacity to measure writing as those that use one skill (independent writing). The underlying skills needed for integrated writing have been found to be complex and go beyond simply avoiding verbatim source use. They require a capacity for discourse synthesis, viewpoint recognition, and connecting one's own argument to the source. Recent rubrics designed in research could facilitate both the teaching of these complex processes and the provision of focused feedback to students in these areas.

5.3.1. *Fairness, justice, and criticality in integrated writing assessment*

Integrated assessment tasks show potential to increase fairness and justice in writing assessment. Research has provided considerable evidence for validity and reliability of these tasks (Cumming et al., 2005; Gebril, 2009, 2010; Kim & Crossley, 2018). However, with a critical lens, new concerns emerge that should be scrutinized. Firstly, providing all test takers with the same content, through reading source material, intends to diminish advantages such as background knowledge or familiarity with genre (Cumming, 2013). Process studies have revealed that writers draw on reading, writing, and the combination of these skills when composing these tasks (Xu & Wu, 2012; Yang & Plakans, 2012). This could make these tasks accessible and fair. However, critical questions remain regarding whether this is indeed the case. Background knowledge has an impact on reading, not just on writing. Thus, if a student or test taker engages with the task and draws on knowledge about the topic for the source text and writing, they would still retain a potential advantage over a test taker without such background. Research should be undertaken to explore this effect in such writing tasks.

A second advantage of these tasks is their attention to context (Huang, 2012; Plakans, 2009). Particularly for academic settings, writing is not done in isolation but with content drawn from reading or listening. Therefore, the task addresses the realities of language use. The 'ecological validity' can increase fairness. However, a lens of criticality might question whether the context of academic writing is itself perpetuating inequity and upholding a dominant variety of writing, that is, one that depends on and cites others' ideas (Hyland, 2009). In fact, there are many ways that different languages, varieties of languages, and cultures incorporate intertextuality. Intertextuality is the concept that all texts are shaped by other texts. An English or 'Western' dominant approach is to cite as an acknowledgement when incorporating ideas or wording published by others. Digging more deeply into intertexuality (Baron, 2019), some might question whether anyone truly 'owns' words or ideas in today's society, where we are constantly reading, talking, and digesting words and ideas of others. Who is the owner? Who is the borrower? Is there a power dimension to this distinction? These questions also recall issues from the previous section's critique of rating and scoring, namely critical questions about standards and bias.

As researchers of integrated writing assessment continue to study these tasks, a critical lens should inform their work on defining the construct, understanding the relationship between reading and writing, developing rubrics and tasks, and exploring underlying values around the ownership of language.

5.4. *Teachers' and learners' knowledge, beliefs, and perspectives on writing assessment*

Research on teachers' and learners' perspectives on writing assessment is integrally linked to classroom use of assessments. Studies in both L1 and L2 writing research have illustrated the importance

https://doi.org/10.1017/S0261444825100876 Published online by Cambridge University Press

of building teachers' skills in assessing writing to boost effective writing instruction. Crusan et al. (2016) investigated the landscape of second language writing teachers' assessment literacy in terms of knowledge, belief, and practices. The majority of respondents to a survey reported learning the theoretical and practical fundamentals of L2 writing assessment through coursework during graduate studies; however, about one-fifth reported they had received no formal education in teaching writing or assessing writing. Some widely used assessment methods, such as portfolio assessment, self-assessment, and scoring rubrics, were favorably received by the majority of the participants and extensively used in their classroom contexts. Noting that L2 writing teachers' beliefs have not been studied extensively, Karaca & Uysal (2021) developed and validated a scale dedicated to measuring L2 writing teachers' beliefs about the teaching and assessment of L2 writing. The strongest belief that the teachers held was about reader-centered writing, closely followed by the important role of motivation in L2 writing. The weakest belief pertained to the role of L1 (i.e. transfer) and discourse-level competency in English writing.

To be self-efficacious, a teacher needs to understand what goes into the preparation of quality writing instruction and how to translate that understanding into practice. Locke and Johnston (2016) defined these as PRE-WRITING INSTRUCTION STRATEGIES and COMPOSITIONAL STRATEGY DEMONSTRATION for L2 writing teachers. They argued for the pivotal role these play in enhancing writing instruction and learning outcomes. For instance, in Dempsey et al. (2009), L1 writing teachers who were aided by an online-based assessment tool in making informed use of an analytic rubric, witnessed significant improvement in their ability to assess student writing. They also reported a heightened degree of self-efficacy in grading learners' writing.

Research has shown that teachers hold a variety of ideas about what constitutes good feedback practice, displayed in varying foci in their commentary to students' writing. For instance, in Marefat and Heydari (2016), the evaluations by 'native-English-speaking' teachers of essays written by Iranian college students were found to be more consistent and reliable, but they tended to be stricter in rating organization. In contrast, 'non-native-English-speaking' teachers' attention was focused more on grammar than on organization. An extensive dataset was analyzed by Dixon and Moxley (2013), involving 118,611 comments made on 17,433 essays written by L1 first-year college students. Instructors primarily drew attention to rhetorical concerns in these essays, such as organization and use of evidence to support arguments. Lower-order concerns of grammar or formatting were pointed out to a lesser degree, reflecting an emphasis on the discourse structure in L1 writing.

A study by Beck et al. (2018) explored the ways in which L2 writing teachers established instructional priorities in response to writing produced by a sample of ninth and tenth grade students. It was found that, in their instructional practice, teachers prioritized readily teachable notions, such as structure, rather than the more abstract cognitive processes that go into writing. In particular, teachers expressed uncertainty and even discomfort about addressing 'knowledge-transformation' (Bereiter & Scardamalia, 1987) aspects of writing, a process in which a writer produces a novel idea from what is presented in the source texts.

To explore teachers' beliefs and perspectives, Ferris (2014) set out to identify the philosophies that teachers uphold as they provide feedback to L2 student writers, and the degree to which their philosophies align with practice in their writing classrooms. She found that currently accepted feedback methods, such as peer review and student-teacher conferences, were favored for use among teachers, and were adopted in line with their philosophies to empower learners as competent writers. A study with a similar focus by Ferris et al. (2011) reported that the teachers were consciously adapting their approach to feedback according to the perceived needs of the students, from explicit attention to grammatical errors to directing students to seek support from external resources, such as a writing center.

Despite these positive findings, research has also shown that the philosophies that teachers express in giving feedback do not always match their practices. Discrepancies have been found between what teachers know or believe they are doing and what they do in the classroom. For example, Li

https://doi.org/10.1017/S0261444825100876 Published online by Cambridge University Press

& Barnard (2011) conducted qualitative interviews with writing tutors working in a university in New Zealand. Despite tutors' commitment to writing improvement, their findings suggested tutors focused on offering feedback directed to justifying more directly the grades assigned to the students rather than feedback supporting the learners' writing development.

An in-depth investigation into the misalignment between beliefs and practices was conducted by Mao and Crosthwaite (2019). Three areas of discrepancies between the teachers' beliefs and teaching practices for giving feedback were investigated. Most teachers stated that they provide more direct than indirect feedback, while, in practice, the opposite appeared to be the case. The teachers also underestimated the number of corrections they provided on local issues, while at the same time overestimating the number of corrections provided on global issues. Niu et al. (2021) compared feedback on essays written by Chinese EFL learners provided by Chinese EFL writing teachers, Chinese peers, and American students. Feedback by the teachers focused more locally on form, whereas the student groups' feedback was more oriented towards meaning. In this study, American students' feedback was observed to bring the highest rates of successful uptake, as opposed to previous research studies which showed a greater uptake for teachers' suggestions.

Research on L2 writing assessment has not paid as much attention to the perspectives and beliefs of L2 test takers. Rather, researchers' attention has primarily shone a spotlight on the written products and their textual features. In some studies, however, learners' perspectives and beliefs about L2 writing assessment has started to be addressed in innovative ways. For example, Aydin (2010) studied 204 learners' perceptions on portfolio assessments through a questionnaire. The students' perceptions were that their proficiency and knowledge improved with a writing portfolio, but some of the processes and procedures in creating them were problematic. In another student-focused study, Kim (2017) investigated Korean EFL students' strategies and challenges with the TOEFL writing section, leading to the conclusion that the score from the TOEFL writing is not an accurate representation of what they know about writing. Kim called for more critical approaches by highlighting test taker's voices about writing tests. Vincelette and Bostic's (2013) centered the perceptions of students and instructors on screencast assessment – an innovative approach to giving feedback that allows for audio and video feedback to writing. Students expressed a strong preference for this method of giving feedback. While it did not lessen the time for instructors' feedback, their comments were more focused on macro-level issues with writing.

Research studies dedicated to uncovering the test taker/learner's perceptions of tests contribute to validity evidence in L2 writing assessment studies. Xie (2015) studied what test takers perceived as critical in their writing to secure high scores on a writing assessment. The researcher administered a survey to 886 college-level Chinese EFL learners and ran an exploratory factor analysis. The researcher identified two different factors underlying the learners' responses. The first factor was the perception that a test taker should use words that are sophisticated and less common as well as more complex sentence structures. The second factor, which was somewhat in conflict with the first, was 'avoiding penalties from raters' by reducing use of unfamiliar lexical and syntactic choices. Taking the latter approach, which is perhaps reactive, was found to be associated more highly with getting higher scores.

In summary, research has revealed complex perspectives of teachers and learners concerning writing assessment in terms of understanding assessment concepts, test-taking strategies, and attitudes. The implications of this research point to the overall importance of teacher writing assessment literacy. This influences the practices and priorities in their writing instruction. Further, research has suggested areas for focus within such literacy, such as understanding the complex and abstract aspects of writing processes and performances. An area in need of further research and scrutiny is students' lived experiences with L2 writing tests and the role that the tests play in mediating opportunities and access, such as success in higher education or employment opportunities (Hamid et al., 2019). Addressing these questions will broaden our understanding of the role and effect of L2 writing tests in teaching and in shaping the lives of language learners.

https://doi.org/10.1017/S0261444825100876 Published online by Cambridge University Press

5.4.1. Research on fairness, justice, and criticality in teachers' and students' beliefs

In research on perspectives, beliefs, and knowledge, teachers have not been asked about issues of fairness, equity, or even ethics in their knowledge of writing assessment in classroom contexts. Are they prepared to recognize and address fairness and justice issues within their practices of assessing writing? Humanizing approaches to research incorporate work with teachers' reflections on their role in systems of inequity as well as experiencing discrimination themselves. In addition to including a critical lens in teachers' professional learning, the field should consider whether access to assessment literacy is equitable. Are some teachers (dis)advantaged based on systems, context, or cultural background in receiving preparation in writing assessment? In Crusan et al. (2016), teachers who were not 'native' English speakers were, in fact, more confident and used more assessment strategies. Others found this group of teachers more consistent and reliable in feedback (Marefat & Heydari, 2016). These findings contradict the value placed on 'native-language' ability, which should not be a proxy to language teacher quality or assessment literacy; in fact, there is counter-evidence to this damaging power dynamic in our field.

Other stakeholders need more attention to address fairness and justice in writing research. Based on our review, the perspectives of learners seems to be a largely under-researched area. However, an encouraging number of researchers have embarked on filling this research gap (Aydin, 2010; Kim, 2017; Vincelette & Bostic, 2013; Xie, 2015), although more is needed. These voices are important to understand the impact that enacting explicit or hidden standards has on how students are marginalized by assessment practices. More work to highlight what students bring to writing classrooms in terms of experience and knowledge of writing and writing assessment could uncover important findings related to fairness. Their backgrounds are critical to their learning from assessments: How do they take feedback from teachers and does the feedback recognize their humanity and empower them? How does the assessment support their identities as writers? How do the interactions around assessment support or disenfranchise learners? These questions should be considered for multilingual writers whose writing may be overwhelmed with error correction. The field should consider how feedback can be effective in improving writing while supporting and encouraging writers.

As with the research on writing assessment in general, a focus on native and non-native teachers of writing emerges in the research on teachers. This is a problematic dichotomy that we must stop relying on to describe teachers or raters because it undervalues their multilingualism and creates a false dichotomy that perpetuates an unequal power dynamic. Noticeably, very little attention has been paid to multilingualism in writing assessment for learning, as part of a teacher's profile, and in classroom contexts.

5.5. Feedback in the writing classroom

Writing assessment in classrooms appears most frequently when teachers provide feedback on students' written work. With this interaction, assessment becomes almost ubiquitous in writing instruction. Its pervasiveness has resulted in a steady line of research to understand teachers' feedback. Corrective feedback (CF) has been an area of attention in writing research for some time, and it can be considered within the context of assessment, teaching, and learning. The previous section described teachers' beliefs and perspectives on feedback, which contributes to the discussion of CF. However, research into responding to L2 students' writing in the classroom has largely concerned itself with identifying the nature of CF by classifying it into different types and comparing their effectiveness in improving learners' subsequent writing. For instance, focused CF has been identified as more effective when it is directed toward specific features of writing, compared to comprehensive but less focused CF. This finding has been substantiated in a study by Bitchener & Knoch (2010), who argued for the effectiveness of focused CF based on its benefits for advanced-level ESL learners in improving their accuracy of article usage, a notoriously challenging aspect of English. Research in the past decade into CF has become more nuanced, with several variations of feedback taking hold as classroom practices, such as dynamic written corrective feedback (Kurzer, 2018). In this mode of feedback, teachers provide specific, targeted, and, most importantly, interactive individualized feedback with a view to helping learners become independent in monitoring and self-editing their written work.

A common concern held by L2 writing teachers in EFL contexts is the possibility that language learners might find processing CF challenging. Zheng & Yu (2018) studied the affective, behavioral, and cognitive engagement of lower proficiency Chinese EFL learners in addressing teacher-written CF. While their affective response to corrective feedback was largely positive, the cognitive and behavioral dimensions did not come up to par with the affective dimension, with little improvement observed in writing accuracy. Waller & Papi (2017) examined the role of language learners' implicit theories of writing intelligence in accepting written corrective feedback (WCF) from others. Implicit theories of intelligence, namely a set of beliefs that the learner holds as to the fixedness or malleability of his or her intelligence, has been proposed to significantly affect motivation. In the study, a questionnaire measuring learners' attitudes towards WCF was analyzed, revealing two major dimensions: FEEDBACK-SEEKING ORIENTATION and FEEDBACK-AVOIDING ORIENTATION. Students who harbored an incremental theory of writing intelligence were more likely to pursue feedback, while those who had a fixed mindset were less open to feedback.

While writing teachers work to provide meaningful feedback, learners are included in this process through peer feedback. Scholarship presents nuanced findings in the benefits of peers' feedback. For example, H. Zhao (2010), in a comparative study, examined learners' use of peer feedback and teacher feedback. The study revealed that, while learners incorporated more teacher feedback in their redrafts than peer feedback, they did so with less understanding of its meaning. Weng et al. (2024) carried out an experimental study to compare the writing scores of two groups which differed by who provided feedback: peers or the teacher. Relative to the control group, students in the experimental condition displayed significantly improvement in appreciation of feedback and appraisal of information. Furthermore, Peng (2024) investigated the differential effects of individual and collaborative processing of teacher feedback on writing development, finding that collaborative processing is associated with greater grammatical accuracy. A case study conducted by Yu & Hu (2017) revealed that the feedback-giving practice is not uniform across learners but differs according to individual factors in relation to the immediate sociocultural context. In their case study, two students were found to focus on different aspects of writing as they provided feedback to peers. One participant paid particular attention to surface-level language in writing, such as lexical choices and grammar, while the other pointed out more global issues of writing such as idea development or content. In their 2025 article, He et al. (2025) suggested that ongoing peer review training deepened Chinese EFL learners' cognitive engagement with peer feedback, revealing asynchronous development between noticing writing problems and understanding peer feedback. This study also demonstrated that the quality of peer feedback improved significantly over the period of training, highlighting the importance of instructing learners on how to provide quality feedback on a peer's writing.

With the advent and development of artificial intelligence, provision of feedback has become computerized. Dikli & Bleyle (2014) compared the perceptions of advanced ESL students about automated feedback and instructor feedback on grammar, usage, and mechanics. The researchers concluded that the students had a sense of trust in the quality of feedback by the automated system, and their positive attitudes towards automated feedback were amplified when it was provided in conjunction with instructor feedback. Xiaosa & Ping (2025) longitudinally examined the affective, behavioral, and cognitive responses of three English language learners to automated feedback, identifying both intra- and inter-individual differences. Engagement with automated feedback depended crucially on the purpose of language learning, the preference of certain language skills over others, the student's language proficiency, and affective factors.

https://doi.org/10.1017/S0261444825100876 Published online by Cambridge University Press

Feedback is the quintessential assessment instrument in teaching writing. The research in this area provides direct connections to teaching and learning in writing instruction. As discussed in the previous section on teacher beliefs and knowledge, the alignment of the teacher's philosophy with student needs is important, including their beliefs about feedback (i.e. method, directness, and focus). This alignment is a critical concern, as teachers' philosophies should also align with the overall course and thus result in useful feedback and valid assessment. Secondly, corrective feedback (CF), has been a major area of interest to the field. Research is fairly conclusive that CF is better when focused, rather than comprehensive. It also suggests that it does not operate independent of context. Interaction in dynamic assessment can facilitate the usefulness of CF. Furthermore, student responsiveness may impact CF, so attention to how they use it and if more support is needed would improve uptake. The majority of studies are cross-sectional (Bitchener, 2012), spanning a relatively short time for data collection and analysis. Considering that language development promoted by feedback requires a long-term developmental trajectory, evidence from more longitudinal studies is important.

Lastly, peer feedback has been an area of focus in the past decade revealing a multitude of factors impacting its success, including the human element on rating/scoring as discussed in the previous section. Having students in the role of giving feedback is also impacted by bias and relationships with peers. Providing ample training and ongoing support in peer feedback could potentially minimize these issues with success.

5.5.1. Fairness, justice, and criticality in feedback research

Similar to the critique of research on features of writing performance, language 'standards' appear in research on feedback explicitly and implicitly. Since feedback is both assessment and instruction, the approach a teacher takes will perpetuate and dominate a classroom. Feedback can support or undermine opportunities to learn, empower, and promote equity in writing instruction. Research has focused on effectiveness or preferences in feedback but not directly taken on the important role that feedback has in assessment and equity. Critical questions include: Does feedback, from teachers or peers, give students equal opportunities to learn and develop their writing? Do teachers include feedback that values students and empowers them as writers? The attention to corrective feedback draws the focus of assessment to correctness and grammatical accuracy (Bitchener & Knoch, 2010), which are based on dominant norms of language standards. In teaching writing, scholars and instructors recognize there is more to writing than lexico-grammatical accuracy, but it remains unclear how accuracy is defined in relation to different varieties of a language.

Another critical aspect of feedback, mentioned in previous sections, is the human factor. Teachers bring prior experience and bias that impacts their feedback and expectations of students. For example, their focus may be on grading rather than learning (Li & Barnard, 2011) or empowerment. The impact of the teacher on feedback is not just an issue of reliability but it can impact fairness in the classroom and justice in terms of access in a learner's future. Strategies have been proposed over the years seeking to ameliorate inequity, such as giving feedback without student names on the submission. However, this is not a problem that is so easily resolved. Rubrics are another attempt to quash inequity in giving feedback; however, this too has flaws as mentioned in the previous section on rating. These are spaces for dominant values to preside unquestioned (teacher bias and rubrics) and also areas that are venues for critical work. Instructor-student negotiated rubrics create opportunities for student input on how and what feedback is given and for peers to bring shared experience to the feedback-giving process. Research on these approaches could inform the potential for feedback to disrupt inequity in writing classrooms. The research on dynamic corrective feedback (Kurzer, 2018) is an example of moving feedback in new directions that could better serve students in light of fairness and justice. Another area of potential research is 'ungrading' which disconnects the feedback-revising loop from any form of points or grades, changing both the value place of 'correctness' and redirecting writing improvement as intrinsic motivation in and of itself.

https://doi.org/10.1017/S0261444825100876 Published online by Cambridge University Press

Table 2. Summary of the themes

Themes	Summary
Features of writing performance	– Research has explored a wide range of textual features in assessments of L2 writing, such as complexity, accuracy, lexis, and fluency (CALF).
	– The scope of investigation has expanded in recent years, with researchers increasingly attending to broader and higher-order constructs, such as functional adequacy, authorial voice, and critical thinking.
	– Recent studies emphasize how writing performance is heavily influenced by the design of assessment tasks.
Rating and scoring L2 writing	– Recent research continues to examine the development, validation, and contextual appropriateness of rating scales used in L2 writing assessment.
	– An increasing number of research studies pay particular attention to the reliability of ratings for L2 writing, using such analytical frameworks and generalizability theory.
Integrated writing assessment	– Process-based studies show that cognitive and composing strategies differ between integrated and independent tasks.
	– Effective integrated writing involves discourse synthesis, self-regulation, and avoiding test-wiseness strategies.
	– Integrated writing is pedagogically valuable, offering students insight into their strengths and areas for improvement.
Teachers' and learners' knowledge and beliefs	– Teachers' assessment literacy is essential for effective writing instruction, but many have limited or no formal training in writing assessment.
	– Learner perspectives are emerging.
	– Teachers often prioritize easily teachable aspects like structure over more abstract cognitive writing processes (e.g. knowledge transformation).
	– Feedback practices are shaped by teacher beliefs but also influenced by grading pressures and institutional norms.
Feedback in the writing classroom	– Corrective and peer feedback research emphasizes type, uptake, and teacher bias.
	– Automated feedback is growing.
	– Dynamic feedback is an effective mechanism to provide input on a student's proficiency in L2 writing.

6. Concluding thoughts on reviewing writing assessment research through a critical lens

Based on our review of over two decades of writing assessment research, we found five major themes explored through research questions related to writing assessment in language teaching and learning: (1) features of writing performance, (2) rating and scoring, (3) integrated assessment, (4) teachers' and learners' knowledge, beliefs, and perspectives, and (5) feedback in the writing classroom. We summarized and highlighted examples of research in each of these areas in relation to and with implications for teaching and learning. For each theme, we provided a critique of the themes related to issues of fairness and justice with a critical lens. Table 2 summarizes findings within each theme of the review.

Layering the framework (Fig. 1) that we proposed over the review of this research, several overarching areas and actions surfaced to enact critical approaches, which we will discuss further as action

https://doi.org/10.1017/S0261444825100876 Published online by Cambridge University Press

items for writing assessment researchers: (1) being more explicit about what/whose standards underlie measures of writing features and definitions of proficiency, (2) not marginalizing multilingualism, and (3) centering voices of stakeholders. While these emerged from our review of a select group articles and journals, we recognize that work which covers these issues may be appearing in venues not explored by this review. Moving forward, we hope the field will continue to investigate and publish research in writing assessment that seeks to dismantle inequity in teaching and learning, as well as to highlight, empower, and value our students' experiences, languages, and writing.

The current approach in researching characteristics of writing at different proficiency levels deserves a critical lens to uncover their embedded norms and values. As Shohamy (1998) states, it is necessary to interrogate 'whose knowledge the tests are based' on (p. 333). A critical approach requires more recognition and transparency on the origins of characteristics such as CALF, or other interpretations of quality and measures used therein. What defines proficiency and proficiency levels in English or any language on which L2 writing is assessed? Questions need to be asked about power, equity, and marginalization in relation to how we characterize writing performances in assessments. These are big questions, each comprising an individual research agenda. Scholars looking at sociocultural features of writing are adding depth to the traditional features that have long been the focus of L2 writing, which is promising (C. G. Zhao, 2013). Ignoring the dominant norms at play in selecting 'features' or defining 'proficiency levels' and not taking them up to trouble our work is perpetuating systems that subjugate multilingual writers.

Scholars in second language acquisition have articulated a 'bilingual turn' in the field (Ortega, 2013), rather than centering theories of language acquisition on monolingualism and sequential language learning (first language + second language). Multilingualism recognizes young learners who develop two or more languages concurrently and adults who use both/multiple languages in writing as resources to think and express themselves (DeCosta et al., 2022). This reality has been addressed in practice through approaches in bilingual education and in the evolving theory of translanguaging (García, 2017; García & Otheguy, 2020; Wei & Lin, 2019). A special issue of *Language Assessment Quarterly* (2019) featured research and practice around multilingual assessment.

More research should delve into the cognitive and socially constructed aspects of multilingualism in writing. This work is foundational in building a construct for multilingual writing. To illustrate, an approach to validity integrated with fairness and justice would insist on centering multilingualism in writing assessment, rather than monolingualism, through a construct that reflects the sophisticated complexity in language use of writers with more than one variety of a language or multiple languages. The challenge of a generalized model of multilingualism will be difficult and imperfect, as scholars who work with the notion of translanguaging (Wei, 2018) emphasize the individualized ways in which people use multiple languages. Therefore, building an assessment that authentically elicits and reflects the way multilinguals use multiple languages is highly complex but not insurmountable. Developing such assessments starts by defining the construct of multilingualism.

For example, rating scales in writing assessments are by design monolingual. An important and intriguing challenge for the field is to examine how scales could address multilingualism. There have been attempts to design multilingual assessments (Guzman-Orth et al., 2019; Lopez, 2023), but when it comes to scoring, challenges emerge, that is, clarity on what the score is telling us about a learners' language and how this information will be used. To value multilingualism in writing assessments, important work is needed to understand the domains of multilingual language use. This relates to performance and processes in composing an assessment. Critical perspectives and multilingual turns are present in language education research (DeCosta et al., 2022; Ortega, 2013), and, thus, writing assessments can draw on this scholarship.

In reviewing writing assessment research, students' voices rarely appear in studies. Engaging more with these stakeholders in research would be a step toward acknowledging power structures in assessment. It is promising to see research that includes the voices of test takers, like that of Xie (2015) who sought to understand their perceptions of writing assessment. Research that attends to both

https://doi.org/10.1017/S0261444825100876 Published online by Cambridge University Press

teachers' voices and agency, as well as scholarship that recognizes and illuminates teacher (and rater) biases, should continue to be undertaken in the field. Related to elevating their voices, how we group and characterize teachers and students also needs attention in writing assessment. For example, the dichotomy of 'native' versus 'non-native' has been repeatedly questioned, and, yet it still routinely appears through the journals reviewed in this time frame. The use of a binary (native/non-native) excludes or diminishes a more common way of languaging by multilinguals. Writing assessment research needs to cease comparing teachers, raters, or students with the imprecise and discriminatory dichotomy of nativeness. Again, pushing back on this is not novel, yet it continues to persist in research questions to categorize writers and teachers. Changing this default in our field will require major rethinking, and yet the reality of multilingualism is not new and is more prevalent than monolingualism. For example, reframing the 'ownership' of English to a Global English approach moves away from the territoriality of language. Shohamy (1998) recognizes that criticality must question complex issues like this in our assessments, as CLT 'perceives of language testing as being caught up in an array of questions concerning educational and social systems' (p. 333).

As we conclude, we point out that, as researchers, we include ourselves in this critique. In much of our past work, especially with integrated writing assessment research, we have not addressed our research through a critical lens. We are working to learn more and consider critical questions of fairness and justice in our research. If this State-of-the-Art article appears as finger pointing, then it is indeed directly pointed at ourselves. Vulnerability and self-reflection should be ongoing in critical work as researchers, providing us with new insights and awareness to inform our work.

In conclusion, the field of writing assessment needs to address questions of fairness, justice, and criticality in our research and practice. In his book, *Antiracism in writing assessment: Teaching and assessing writing for a socially just future*, Inoue (2015) states:

> Understanding classroom writing assessment as an ecology that can be designed and cultivated shows that the assessment of writing in not simply a decision about whether to use a portfolio or not, or what rubric to use. It is about cultivating and nurturing complex systems that are centrally about sustaining fairness and diverse complexity (p. 12).

This is essential in providing equity for multilingual language learners in schools and classrooms. A critical perspective includes acknowledging that power dimensions intersect and underwrite educational policy, assessments, and instruction (Flores & Rosa, 2015; Kubota, 2020).

Writing assessments have the potential to value multiple languages and multilingualism as well as provide opportunities; however, they can also have the opposite effect, resulting in denied access. In our review of research over the past decade, the field of L2 writing assessment appears to lack a strong critical stance in research. We need to unpack how our writing assessments are perpetuating hegemony and hierarchies in order to dismantle systems that suppress rather than uplift language learners. The scholars in language testing calling for critical language testing (Lynch, 2001; Shohamy, 1998, 2001) and taking social perspectives (McNamara & Roever, 2006; Roever & Wigglesworth, 2019; Schissel & Khan, 2021) have provided directions for this work, which could be extended by drawing on theory such as CritLing and critical discourse analysis (Catalano & Waugh, 2020), critical language awareness (Britton & Leonard, 2020), critical race theory (Delgado & Stanfincic, 2017) or raciolinguistics (Degollado, 2019; Flores & Rosa, 2015). These approaches can illuminate and respond to power, linguicism, racism, and inequity in second language writing assessment.

7. Questions arising

In this review, we found established themes in research on writing assessment in published research over the past 20 years in a select group of journals in the field. Through this work we learned about writing of language learners and their teachers as well as the instruments and characteristics of writing in assessments. Reviewing this research through a framework of fairness and justice (Fig. 1),

https://doi.org/10.1017/S0261444825100876 Published online by Cambridge University Press

with attention to critical inquiry, we find considerable room for work to address inequities, power dimensions, and marginalized voices in our field. Research plays an important role in uncovering and unpacking assumptions. It is also necessary in rebuilding and identifying pathways forward. From our review of literature and critique, we propose a list of questions for future research to consider. These questions are a potential starting point and not a comprehensive list. Some directly address individual themes in the review, while others can be applied to research across several themes.

- Whose language is the 'standard' for language proficiency in writing assessment? Why?
- What domains are privileged in our writing assessments?
- What practices can best illuminate rater and test user bias in writing assessment?
- How can teachers reflect on the role of assessment in marginalizing students' languages?
- How can writing assessment avoid prescriptive approaches to teaching and learning?
- What constructs capture multilingual writing ability?
- How can we disrupt the false dichotomy of 'native' and 'non-native' speaker?
- What are ways to include students in writing assessment research?
- How can test development enable fair and just writing assessment?

Reflecting on and researching these questions can transform writing assessment in teaching and learning to create more equitable opportunities for learners and their languages. The questions also complicate processes and create messiness that is difficult to navigate in assessment, which has traditionally upheld precision and accuracy as necessary qualities. However, using a framework that situates validity with fairness and justice (McNamara & Ryan, 2011) within the landscape of context and with attention to criticality can create space for writing assessment research to recognize and repel inequity and to value the diversity of languages and languaging of our learners (Lynch, 2001; Shohamy, 1998, 2001).

References

References with * notation were in the sample reviewed for themes.

Adams, R., Nik Mohd Alwi, N. A., & Newton, J. (2015). Task complexity effects on the complexity and accuracy of writing via text chat. *Journal of Second Language Writing*, *29*, 64–81. doi:10.1016/j.jslw.2015.06.002*

Asencion, Y. (2008). Investigating the reading-to-write construct. *Journal of English for Academic Purposes*, *7*(3), 140–150. doi:10.1016/j.jeap.2008.04.001*

Attali, Y. (2016). A comparison of newly-trained and experienced raters on a standardized writing assessment. *Language Testing*, *33*(1), 99–115. doi:10.1177/0265532215582283*

Attali, Y., Lewis, W., & Steier, M. (2013). Scoring with the computer: Alternative procedures for improving the reliability of holistic essay scoring. *Language Testing*, *30*(1), 125–141. doi:10.1177/0265532212452396*

Aydin, S. (2010). EFL writers' perceptions of portfolio keeping. *Assessing Writing*, *15*(3), 194–203. doi:10.1016/j.asw.2010.08.001*

Banerjee, J., Yan, X., Chapman, M., & Elliott, H. (2015). Keeping up with the times: Revising and refreshing a rating scale. *Assessing Writing*, *26*, 5–19. doi:10.1016/j.asw.2015.07.001*

Barkaoui, K. (2010). Explaining ESL essay holistic scores: A multilevel modeling approach. *Language Testing*, *27*(4), 515–535. doi:10.1177/0265532210368717*

Baron, S. (2019). *The birth of intertextuality: The riddle of creativity*. Routledge.

Beck, S. W., Llosa, L., Black, K., & Anderson, A. T. G. (2018). From assessing to teaching writing: What teachers prioritize. *Assessing Writing*, *37*, 68–77. doi:10.1016/j.asw.2018.03.003*

Behizadeh, N., & Engelhard, G. (2014). Development and validation of a scale to measure perceived authenticity in writing. *Assessing Writing*, *21*, 18–36. doi:10.1016/j.asw.2014.02.001*

Bejar, I. I., Flor, M., Futagi, Y., & Ramineni, C. (2014). On the vulnerability of automated scoring to construct-irrelevant response strategies (CIRS): An illustration. *Assessing Writing*, *22*, 48–59. doi:10.1016/j.asw.2014.06.001*

Bereiter, C., & Scardamalia, M. (1987). *The psychology of written composition*. L. Erlbaum.

Bitchener, J. (2012). Written corrective feedback for L2 development: Current knowledge and future research. *TESOL Quarterly*, *46*(4), 855–860. doi:10.1002/tesq.62*

Bitchener, J., & Knoch, U. (2010). Raising the linguistic accuracy level of advanced L2 writers with written corrective feedback. *Journal of Second Language Writing*, *19*(4), 207–217. doi:10.1016/j.jslw.2010.10.002*

Bouwer, R., Béguin, A., Sanders, T., & van den Bergh, H. (2015). Effect of genre on the generalizability of writing scores. *Language Testing*, *32*(1), 83–100. doi:10.1177/0265532214542994*

Britton, E. R., & Leonard, R. L. (2020). The social justice potential of critical reflection and critical language pedagogies for L2 writers. *Journal of Second Language Writing*, *50*, 10076. doi:10.1016/j.jslw.2020.100776

Brown, K. D. (2013). Trouble on my mind: Toward a framework of humanizing critical sociocultural knowledge for teaching and teacher education. *Race Ethnicity and Education*, *16*(3), 316–338. doi:10.1080/13613324.2012.725039

Catalano, T., & Waugh, L. R. (2020). *Critical discourse analysis, critical discourse studies, and beyond*. Springer.

Chan, S., Inoue, C., & Taylor, L. (2015). Developing rubrics to assess the reading-into-writing skills: A case study. *Assessing Writing*, *26*, 20–37. doi:10.1016/j.asw.2015.07.004*

Cohen, A. D. (1998). Strategies and processes in test taking and SLA. In M. H. Long & J. C. Richards (Eds.), *Interfaces between second language acquisition and language testingresearch* (pp. 90–111). Cambridge University Press.

Crusan, D. (2010). *Assessment in the second language writing classroom*. University of Michigan Press.

Crusan, D., Plakans, L., & Gebril, A. (2016). Writing assessment literacy: Surveying second language teachers' knowledge, beliefs, and practices. *Assessing Writing*, *28*, 43–56. doi:10.1016/j.asw.2016.03.001*

Cumming, A. (2013). Assessing integrated writing tasks for academic purposes: Promises and perils. *Language Assessment Quarterly*, *10*(1), 1–8. doi:10.1080/15434303.2011.622016*

Cumming, A., Kantor, R., Baba, K., Erdosy, U., Eouanzoui, K., & James, M. (2005). Differences in written discourse in independent and integrated prototype tasks for next generation TOEFL. *Assessing Writing*, *10*(1), 5–43. doi:10.1016/j.asw.2005.02.001*

De Costa, P. I. (2018). Toward greater diversity and social equality in language education research. *Critical Inquiry in Language Studies*, *15*(4), 302–307. doi:10.1080/15427587.2018.1443267

DeCosta, P., Li, W., & Lee, J. (Eds.). (2022). *International students' multilingual literacy practices: An assest-based approach to understanding academic discourse socialization*. Mulitlingual Matters.

Degollado, E. D. (2019). *The storied lives of fronterize bilingual maestras: Constructing language and literacy ideologies in neplanta* [Unpublished doctoral dissertation]. University of Texas at Austin.

Delgado, R., & Stanfincic, J. (2017). *Critical race theory: An introduction.* (3rd ed.) NYU Press.

Dempsey, M. S., PytlikZillig, L. M., & Bruning, R. H. (2009). Helping preservice teachers learn to assess writing: Practice and feedback in a Web-based environment. *Assessing Writing*, *14*(1), 38–61. doi:10.1016/j.asw.2008.12.003*

Dikli, S., & Bleyle, S. (2014). Automated essay scoring feedback for second language writers: How does it compare to instructor feedback? *Assessing Writing*, *22*, 1–17. doi:10.1016/j.asw.2014.03.006*

Dixon, Z., & Moxley, J. (2013). Everything is illuminated: What big data can tell us about teacher commentary. *Assessing Writing*, *18*(4), 241–256. doi:10.1016/j.asw.2013.08.002*

Ferris, D., Brown, J., Liu, H. S., & Stine, M. E. A. (2011). Responding to L2 students in college writing classes: Teacher perspectives. *TESOL Quarterly*, *45*(2), 207–234. doi: 10.5054/tq.2011.247706*

Ferris, D. R. (2014). Responding to student writing: Teachers' philosophies and practices. *Assessing Writing*, *19*, 6–23. doi:10.1016/j.asw.2013.09.004*

Flores, N., & Rosa, J. (2015). Undoing appropriateness: Raciolinguistic ideologies and language diversity in education. *Harvard Educational Review*, *85*(2), 149–171. doi:10.17763/0017-8055.85.2.149

Frear, M. W., & Bitchener, J. (2015). The effects of cognitive task complexity on writing complexity. *Journal of Second Language Writing*, *30*, 45–57. doi:10.1016/j.jslw.2015.08.009*

García, O. (2017). Translanguaging in schools: Subiendo y bajando, bajando y subiendo as afterword. *Journal of Language, Identity, and Education*, *4*(4), 256–263. doi:10.1080/15348458.2017.1329657

García, O., & Otheguy, R. (2020). Plurilingualim and translanguaging: Commonalities and divergences. *International Journal of Bilingual Education and Bilingualism*, *23*(1), 17–35. doi:10.1080/13670050.2019.1598932

Gebril, A. (2009). Score generalizability of academic writing tasks: Does one test method fit it all? *Language Testing*, *26*(4), 507–531. doi:10.1177/0265532209340188*

Gebril, A. (2010). Bringing reading-to-write and writing-only assessment tasks together: A generalizability analysis. *Assessing Writing*, *15*(2), 100–117. doi:10.1016/j.asw.2010.05.002*

Gebril, A., & Brown, G. T. (2019). A learning-oriented assessment perspective. In J. de Dios Martínez Agudo (Ed.), *Quality in TESOL and teacher education: From a results culture towards a quality culture*. Routledge.

Gebril, A., & Plakans, L. (2014). Assembling validity evidence for assessing academic writing: Rater reactions to integrated tasks. *Assessing Writing*, *21*, 56–73. doi:10.1016/j.asw.2014.03.002*

Grice, H. P. (1975). Logic and conversation. Cole, Peter, and Morgan, Jerry L. Eds., *Speech acts* 3 Syntax and semantics, 41–58. Brill.

Guzman-Orth, D. A., Lopez, A. A., & Tolentino, F. (2019). Exploring the use of a dual language assessment task to assess young English learners. *Language Assessment Quarterly*, *16*(4–5), 447–463. doi:10.1080/15434303.2019.1674314

Hamid, M. O., Hardy, I., & Reyes, V. (2019). Test-takers' perspectives on a global test of English: Questions of fairness, justice, and validity. *Language Testing in Asia*, *9*(1). doi:10.1186/s40468-019-0092-9*

Hamp-Lyons, L. (2002). The scope of writing assessment. *Assessing Writing*, *8*, 5–16. doi:10.1016/s1075-2935(02)00029-6*

https://doi.org/10.1017/S0261444825100876 Published online by Cambridge University Press

Han, C. (2019). A generalizability theory study of optimal measurement design for a summative assessment of English/Chinese consecutive interpreting. *Language Testing*, *36*(3), 419–438. doi:10.1177/0265532218809396*

Hannah, L., Jang, E. E., Shah, M., & Gupta, V. (2023). Validity arguments for automated essay scoring of young students' writing traits. *Language Assessment Quarterly*, *20*(4–5), 399–420. doi:10.1080/15434303.2023.2288253*

He, J., Xia, J., Zhang, C. M., & Liu, J. N. (2025). Promoting cognitive engagement with peer feedback through peer review training: The case of Chinese tertiary-level EFL learners. *Assessing Writing*, *65*, 100947. doi:10.1016/j.asw.2025.100947*

Huang, J. (2012). Using generalizability theory to examine the accuracy and validity of large-scale ESL writing assessment. *Assessing Writing*, *17*(3), 123–139. doi:10.1016/j.asw.2011.12.003*

Hyland, K. (2009). *Teaching and researching writing* (2nd ed.). Longman.

Inbar-Lourie, O. (2008). Constructing a language assessment knowledge base: A focus on language assessment courses. *Language Testing*, *25*(3), 385–402. doi: 10.1177/0265532208090158

Inoue, A. B. (2015). *Antiracist writing assessment ecologies*. Parlor Press.

Janssen, G., Meier, V., & Trace, J. (2015). Building a better rubric: Mixed methods rubric revision. *Assessing Writing*, *26*, 51–66. doi:10.1016/j.asw.2015.07.002*

Jones, N., & Saville, N. (2016). *Learning oriented assesment: A systemic approach*, Cambridge University Press.

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

Karaca, M., & Uysal, H. H. (2021). The development and validation of an inventory on English writing teacher beliefs. *Assessing Writing*, *47*, 100507. doi:10.1016/j.asw.2020.100507*

Kim, E.-Y. J. (2017). The TOEFL iBT writing: Korean students' perceptions of the TOEFL iBT writing test. *Assessing Writing*, *33*, 1–11. doi:10.1016/j.asw.2017.02.001*

Kim, M., & Crossley, S. A. (2018). Modeling second language writing quality: A structural equation investigation of lexical, syntactic, and cohesive features in source-based and independent writing. *Assessing Writing*, *37*, 39–56. doi:10.1016/j.asw.2018.03.002*

Knoch, U. (2011). Rating scales for diagnostic assessment of writing: What should they look like and where should the criteria come from? *Assessing Writing*, *16*(2), 81–96. doi:10.1016/j.asw.2011.02.003*

Knoch, U., & Sitajalabhorn, W. (2013). A closer look at integrated writing tasks: Towards a more focused definition for assessment purposes. *Assessing Writing*, *18*(4), 300–308. doi:10.1016/j.asw.2013.09.003*

Kubota, R. (2020). Confronting epistemological racism, decolonizing scholarly knowledge: Race and gender in applied linguistics. *Applied Linguistics*, *42*(5), 712–732. doi:10.1093/applin/amz033

Kubota, R., & Miller, E. R. (2017). Re-examining and re-envisioning criticality in language studies: Theories and praxis. *Critical Inquiry in Language Studies*, *12*(2–3), 129–157. doi:10.1080/15427587.2017.1290500

Kuiken, F., & Vedder, I. (2017). Functional adequacy in L2 writing: Towards a new rating scale. *Language Testing*, *34*(3), 321–336. doi:10.1177/0265532216663991*

Kumashiro, K. K. (2000). Teaching and learning through desire, crisis, and difference: Perverted reflections on anti-oppressive education. *The Radical Teacher*, *58*, 6–11.

Kunnan, A. J. (2014). Fairness and justice in language assessment: Principles and public reasoning. *Alternative pedagogies in the English language and communication classroom*, 36-39.

Kurzer, K. (2018). Dynamic written corrective feedback in developmental multilingual writing classes. *TESOL Quarterly*, *52*(1), 5–33. doi:10.1002/tesq.366*

Kyle, K. (2020). The relationship between features of source text use and integrated writing quality. *Assessing Writing*, 45, 100467. doi:10.1016/j.asw.2020.100467

Lallmamode, S. P., Mat Daud, N., & Abu Kassim, N. L. (2016). Development and initial argument-based validation of a scoring rubric used in the assessment of L2 writing electronic portfolios. *Assessing Writing*, *30*, 44–62. doi:10.1016/j.asw.2016.06.001*

Lee, K., Liao, R. J., Hsiao, I. C. V., Park, J., & Ye, Y. (2025). Predicting inappropriate source use from scores of language use, source comprehension, and organizational features: A study using generalized linear models. *Assessing Writing*, *64*, 100934. doi:10.1016/j.asw.2025.100934*

Li, J., & Barnard, R. (2011). Academic tutors' beliefs about and practices of giving feedback on students' written assignments: A New Zealand case study. *Assessing Writing*, *16*(2), 137–148. doi:10.1016/j.asw.2011.02.004*

Lim, G. S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing*, *28*(4), 543–560. doi:10.1177/0265532211406422*

Ling, G. (2017). Are TOEFL iBT® writing test scores related to keyboard type? A survey of keyboard-related practices at testing centers. *Assessing Writing*, 31, 1–12. doi: 10.1016/j.asw.2016.04.001

Locke, T., & Johnston, M. (2016). Developing an individual and collective self-efficacy scale for the teaching of writing in high schools. *Assessing Writing*, *28*, 1–14. doi:10.1016/j.asw.2016.01.001*

Lopez, A. A. (2023). Enabling multilingual practices in English language proficiency assessments for young learners. In Raza, K., Reynolds, D., & Coombe, C. (Eds)., *Handbook of multilingual TESOL in practice*. (pp. 359–371). Springer Nature Singapore.

https://doi.org/10.1017/S0261444825100876 Published online by Cambridge University Press

Lynch, B. (2001). Rethinking assessment from a critical perspective. *Language Testing*, *18*(4), 351–372. doi:10.1177/026553220101800403

Mao, S., & Crosthwaite, P. (2019). Investigating written corrective feedback: (Mis)alignment of teachers' beliefs and practices. *Journal of Second Language Writing*, *45*, 46–60. doi:10.1016/j.jslw.2019.05.004*

Marefat, F., & Heydari, M. (2016). Native and Iranian teachers' perceptions and evaluation of Iranian students' English essays. *Assessing Writing*, *27*, 24–36. doi:10.1016/j.asw.2015.10.001*

McCurry, D. (2010). Can machine scoring deal with broad and open writing tests as well as human readers? *Assessing Writing*, *15*(2), 118–129. doi:10.1016/j.asw.2010.04.002*

McNamara, T., & Knoch, U. (2019). *Fairness, justice, and language assessment*. Oxford University Press.

McNamara, T., & Roever, C. (2006). *Language testing: The social dimension*. Blackwell.

McNamara, T., & Ryan, K. (2011). Fairness versus justice in language testing: The place of English literacy in the Australian citizenship test. *Language Assessment Quarterly*, *8*(2), 161–178. doi:10.1080/15434303.2011.565438

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13–103). Macmillan.

Michel, M., Révész, A., Lu, X., Kourtali, N. E., Lee, M., & Borges, L. (2020). Investigating L2 writing processes across independent and integrated tasks: A mixed-methods study. *Second Language Research*, *36*(3), 307–334. doi:10.1177/0267658320915501*

Niu, R., Shan, P., & You, X. (2021). Complementation of multiple sources of feedback in EFL learners' writing. *Assessing Writing*, *49*, 100549. doi:10.1016/j.asw.2021.100549*

Ohta, R., Plakans, L. M., & Gebril, A. (2018). Integrated writing scores based on holistic and multi-trait scales: A generalizability analysis. *Assessing Writing*, *38*, 21–36. doi:10.1016/j.asw.2018.08.001*

Ong, J., & Zhang, L. J. (2010). Effects of task complexity on the fluency and lexical complexity in EFL students' argumentative writing. *Journal of Second Language Writing*, *19*(4), 218–233. doi:10.1016/j.jslw.2010.10.003*

Ortega, L. (2013). SLA for the 21st century: Disciplinary progress, transdisciplinary relevance, and the bi/multilingual turn. *Language Learning*, *63*(s1), 1–24. doi:10.1111/j.1467-9922.2012.00735.x

Peng, C. X. (2024). Beyond accuracy gains: Investigating the impact of individual and collaborative feedback processing on L2 writing development. *Assessing Writing*, *61*, 100876. doi:10.1016/j.asw.2024.100876*

Plakans, L. (2009). Discourse synthesis in integrated second language writing assessment. *Language Testing*, *26*(4), 561–587. doi:10.1177/0265532209340192*

Plakans, L., & Gebril, A. (2013). Using multiple texts in an integrated writing assessment: Source text use as a predictor of score. *Journal of Second Language Writing*, *22*(3), 217–230. doi:10.1016/j.jslw.2013.02.003*

Plakans, L., Gebril, A., & Bilki, Z. (2019). Shaping a score: The impact of fluency, accuracy, and complexity on integrated skills performances. *Journal of Language Testing*, *36*(2), 161–179. doi:10.1177/0265532216669537*

Poe, M., & Elliot, N. (2019). Evidence of fairness: Twenty-five years of research in assessing writing. *Assessing Writing*, *42*, 2–21. doi:10.1016/j.asw.2019.100418

Potter, A., Shortt, M., Goldshtein, M., & Roscoe, D. R. (2025). Assessing academic language in tenth grade essays using natural language processing. *Assessing Writing*, 64, 100921. doi:10.1016/j.asw.2025.100921

Ramineni, C. (2013). Validating automated essay scoring for online writing placement. *Assessing Writing*, *18*(1), 40–61. doi:10.1016/j.asw.2012.10.005*

Randall, J. (2021). 'Color neutral' is not a thing: Redefining construct definition and representation through a justice-oriented critical anti-racist lens. *Educational Measurement: Issues and Practice*, *4*(4), 82–90. doi:10.1111/emip.12429

Roever, C., & Wigglesworth, G. (2019). Social perspectives on language testing: Papers in honour of Tim McNamara. Peter Lang.

Saldaña, J. (2011). *Fundamentals of qualitative research*. Oxford University Press.

Sato, T. (2024). 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. doi:10.1177/02655322231190058*

Saxton, E., Belanger, S., & Becker, W. (2012). The critical thinking analytic rubric (CTAR): Investigating intra-rater and inter-rater reliability of a scoring mechanism for critical thinking performance assessments. *Assessing Writing*, *17*(4), 251–270. doi:10.1016/j.asw.2012.07.002*

Schissel, J. L., & Khan, K. (2021). Responsibilities and opportunities in language testing with respect to historicized forms of socio-political discrimination: A matter of academic citizenship. *Language Testing*, 1–9. doi:10.1177/02655322211028590

Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modeling. *Language Testing*, *22*(1), 1–30. doi:10.1191/0265532205lt295oa.

Sealey-Ruiz, Y. (2021). Racial literacy. A policy research brief. *National Council of Teachers of English.*

Shepard, L. A. (2016). Testing and assessment for the good of education: Contributions of AERA Presidents, 1915–2015. *Educational Researcher*, *45*(2), 112–121. doi:10.3102/0013189x16639599

Shohamy, E. (1998). Critical language testing and beyond. *Studies in Educational Evaluation*, *24*(4), 331–345. doi:10.1016/s0191-491x(98)00020-0

Shohamy, E. (2001). *The power of tests: A critical perspective on the uses of language tests*. Longman.

https://doi.org/10.1017/S0261444825100876 Published online by Cambridge University Press

Sickinger, R., Brunfaut, T., & Pill, J. (2025). Comparative judgement for evaluating young learners' EFL writing performances: Reliability and teacher perceptions of holistic and dimension-based judgements. *Language Testing*, *42*(2), 137–166. doi:10.1177/02655322241288847*

Soto, C. (2022). Language, pedagogy, and discourses of criticality in late capitalism. *Applied Linguistics and Politics*, *235*. doi:10.5040/9781350098268.ch-010

Spivey, N. N. (1984). *Discourse synthesis: Constructing texts in reading and writing*. International Reading Association.

Taguchi, N., Crawford, W., & Wetzel, D. Z. (2013). What linguistic features are indicative of writing quality? A case of argumentative essays in a college composition program. *TESOL Quarterly*, *47*(2), 420–430. doi:10.1002/tesq.91*

Trites, L., & McGroarty, M. (2005). Reading to learn and reading to integrate: New tasks for reading comprehension tests? *Language Testing*, *22*(2), 174–210. doi:10.1191/0265532205lt299oa*

Turner, C. E., & Purpura, J. E. (2016). Learning-oriented assessment in second and foreign language classrooms. In D. Tsagari & J. Baneerjee (Eds.), *Handbook of second language Assessment* (pp. 255–272). De Gruyter.

Tywoniw, R., & Crossley, S. (2019). The effect of cohesive features in integrated and independent L2 writing quality and text classification. *Language Education and Assessment*, *2*(3), 110–134. doi:10.29140/lea.v2n3.151*

Uludag, P., Lindberg, R., McDonough, K., & Payant, C. (2019). Exploring L2 writers' source-text use in an integrated writing assessment. *Journal of Second Language Writing*, *46*, 100670. doi:10.1016/j.jslw.2019.100670*

Vincelette, E. J., & Bostic, T. (2013). Show and tell: Student and instructor perceptions of screencast assessment. *Assessing Writing*, *18*(4), 257–277. doi:10.1016/j.asw.2013.08.001*

Waller, L., & Papi, M. (2017). Motivation and feedback: How implicit theories of intelligence predict L2 writers' motivation and feedback orientation. *Journal of Second Language Writing*, *35*, 54–65. doi:10.1016/j.jslw.2017.01.004*

Watanabe, Y. (2001). *Read-to-write tasks for the assessment of second language academic writing skills: Investigating text features and rater reactions* [Unpublished doctoral dissertation] University of Hawaii.

Wei, L. (2018). Translanguaging as a practical theory of language. *Applied Linguistics*, *39*(1), 9–30. doi:10.1093/applin/amx039

Wei, L., & Lin, A. M. (2019). Translanguaging classroom discourse: Pushing limits, breaking boundaries. *Classroom Discourse*, *10*(3–4), 209–215. doi:10.1080/19463014.2019.1635032.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, *15*(2), 263–287. doi:10.1177/026553229801500205*

Weigle, S. C. (2002). *Assessing writing*. Cambridge University Press.

Weigle, S. C., & Parker, K. (2012). Source text borrowing in an integrated reading/writing assessment. *Journal of Second Language Writing*, *21*(2), 118–133. doi:10.1016/j.jslw.2012.03.004*

Weng, F., Zhao, C. G., & Chen, S. (2024). Effects of peer feedback in English writing classes on EFL students' writing feedback literacy. *Assessing Writing*, *61*, 100874. doi:10.1016/j.asw.2024.100874*

White, E. (2019). (Re)visiting 25 year of writing assessment. *Assessing Writing*, *42*, 1–6. doi:10.1016/j.asw.2019.100419

Winke, P., & Lim, H. (2015). ESL essay raters' cognitive processes in applying the Jacobs et al. rubric: An eye-movement study. *Assessing Writing*, *25*, 38–54. doi:10.1016/j.asw.2015.05.002*

Wolfe-Quintero, K., Inagaki, S., & Kim, H.-Y. (1998). *Second language development in writing: Measures of fluency, accuracy, and complexity*. University of Hawai'i, Second Language Teaching and Curriculum Center.

Xiaosa, L., & Ping, K. (2025). How L2 student writers engage with automated feedback: A longitudinal perspective. *Assessing Writing*, *64*, 100919. doi:10.1016/j.asw.2025.100919*

Xie, Q. (2015). 'I must impress the raters!' An investigation of Chinese test-takers' strategies to manage rater impressions. *Assessing Writing*, *25*, 22–37. doi:10.1016/j.asw.2015.05.001*

Xu, Y., & Wu, Z. (2012). Test-taking strategies for a high-stakes writing test: An exploratory study of 12 Chinese EFL learners. *Assessing Writing*, *17*(3), 174–190. doi:10.1016/j.asw.2012.03.001*

Yancey, K. B. (1999). Looking back as we look forward: Historicizing writing assessment. *College Composition and Communication*, *50*(3), 483–503. doi:10.58680/ccc19991341

Yang, H.-C., & Plakans, L. (2012). Second language writers' strategy use and performance on an integrated reading-listening-writing task. *TESOL Quarterly*, *46*(1), 80–103. doi:10.1002/tesq.6*

Yu, S., & Hu, G. (2017). Understanding university students' peer feedback practices in EFL writing: Insights from a case study. *Assessing Writing*, *33*, 25–35. doi:10.1016/j.asw.2017.03.004*

Zhang, Y., & Ouyang, J. (2023). Linguistic complexity as the predictor of EFL independent and integrated writing quality. *Assessing Writing*, *56*, 100727. doi:10.1016/j.asw.2023.100727*

Zhao, C. G. (2013). Measuring authorial voice strength in L2 argumentative writing: The development and validation of an analytic rubric. *Language Testing*, *30*(2), 201–230. doi:10.1177/0265532212456965*

Zhao, H. (2010). Investigating learners' use and understanding of peer and teacher feedback on writing: A comparative study in a Chinese English writing classroom. *Assessing Writing*, *15*(1), 3–17. doi:10.1016/j.asw.2010.01.002*

Zheng, Y., & Yu, S. (2018). Student engagement with teacher written corrective feedback in EFL writing: A case study of Chinese lower-proficiency students. *Assessing Writing*, *37*, 13–24. doi:10.1016/j.asw.2018.03.001*

https://doi.org/10.1017/S0261444825100876 Published online by Cambridge University Press

Endnotes

[1] Our review primarily focuses on research related to writing assessments in an additional (L2) language; however, some studies of L1 writing were kept in the pool if they included multilingual writers or seemed relevant to this population.

Lia Plakans is a Professor of Multilingual Education in the Department of Teaching and Learning at the University of Iowa's College of Education. Her teaching and research focus on second language assessment, literacy, and language education, with particular emphasis on the integration of reading and writing, assessment design, and multilingual learners' experiences. She has co-authored books including *Assessment myths: Applying second language research to classroom teaching* (University of Michigan Press, 2015) and *Reading and writing for academic success* (University of Michigan Press, 2003). In her scholarship, she seeks to contribute to equity-centered educator preparation and assessment reform. lia-plakans@uiowa.edu

Kwangmin Lee is an Assistant Professor of Teaching English to Speakers of Other Languages (TESOL) in the Department of Special Education and Literacy Studies at Western Michigan University. He works with in-service language educators across Michigan as they pursue a master's degree in TESOL. His research centers on second and foreign language assessment, particularly integrated writing assessment and quantitative research methods. Most recently, he has been expanding his research repertoire by exploring the use of Artificial Intelligence (AI) and Machine Learning (ML) in language assessment to support task innovation. kwangmin.1.lee@wmich.edu

Cite this article: Plakans, L., & Lee, K. (2025). Fairness, justice, and criticality: Reviewing second language writing assessment. *Language Teaching*, 1–28. https://doi.org/10.1017/S0261444825100876

https://doi.org/10.1017/S0261444825100876 Published online by Cambridge University Press