Macrothink Institute™

Journal of Educational Issues
ISSN 2377-2263
2026, Vol. 12, No. 2

A Review of Strategies to Improve Rater Consistency in EFL Writing Assessment

Brandon Baumeister, MSEd

Montgomery Preparatory Academy

2745 S Smithville Rd, Dayton, OH 45420, USA

Tel: 1-937-980-3411 E-mail: brandonbaumeister94715@gmail.com

Darien Wilkerson, M.S, LPC, NCC

Counseling and Psychological Services, University of Michigan-Flint

303 E. Kearsley Street, Flint, MI, 48502, USA

Tel: 1-810-762-3300 E-mail: wilkers@umich.edu

Received: July 21, 2026 Accepted: October 8, 2026 Published: November 18, 2026

doi:10.5296/jei.v12i2.23936 URL: https://doi.org/10.5296/jei.v12i2.23936

Abstract

As interest in second language writing assessment has grown, the role of raters has come under increasing scrutiny. Although raters aim to score fairly, research consistently shows variation in severity and scoring patterns. This narrative literature review synthesizes studies on rater bias in EFL writing assessment and evaluates strategies for improving consistency. Five main approaches are considered: rater training, using multiple raters, avoiding rater fatigue, rubric development, and Multi-Faceted Rasch Measurement (MFRM). Each is discussed in terms of its benefits, limitations, and practical applications. The paper also distinguishes between intra-rater and inter-rater consistency, examines how bias appears in classroom, institutional, and large-scale testing contexts, and considers the broader implications for teachers, institutions, and test developers. The review concludes that no single strategy can eliminate inconsistency, but combining approaches offers the most effective path toward reducing bias and promoting fairness in EFL writing assessment.

Keywords: Rater bias, Rater consistency, EFL writing assessment

1. Introduction

1.1 Introduce the Problem

In EFL instruction, writing is an important skill as it not only demonstrates a student's discourse and rhetorical abilities, but also their proficiency in mechanical skills such as grammar and vocabulary (Sundeen, 2014). An instructor's method of assessment can directly impact the improvement of a student's writing skills (Klenowski, 2009). Due to the high stakes and significant consequences involved in assessment, it is "very important to follow rigorous standards to guarantee that our assessments are valid, fair, and appropriate to the context and purposes for which they are designed" (Crusan, 2010, p. 32). Katz (2014) states that the assessment of students' writing is a central responsibility of teachers and should be closely linked to syllabus design, lesson planning, task and assessment development, and feedback processes. Designing and assessing an effective writing assignment requires "much more than simply thinking of a topic for test takers to write about and then using our own judgement to rank order the resulting writing samples" (Weigle, 2002, p. 2).

Writing remains a difficult skill to assess and many instructors try to avoid assessing writing altogether. Instructors often claim that they are unsure about what constitutes a good and fair writing assessment (Weigle, 2007). Zhang (2012) analyzed college-level EFL written instruction and found that instructors adopted an ambiguous assessment system, and subjective factors largely affected the assessment. As a result, written assessment is often not systematic and reflects rater bias.

Bias is defined as a systematic error that disadvantages the test performance of one group (Shepard et al., 1981). Central to bias analysis is studying unexpected interactions between rater judgement and test takers' performance. Studies have found significant differences in rater severity attributed to bias interactions (Du et al., 1996). Bias can appear in many aspects of the assessment process. For example, McNamara (1996) found that raters assessing the Occupational English Test (OET) in Australia were highly influenced by the grammatical accuracy of the test takers despite the test being communicative in nature. The raters were unaware that they had such bias. Du et al. (1996) in a study of a middle and high school writing assessment program, investigated sources of bias based on ethnic group, gender, and topic type.

They found that significant rater bias accounted for 11% of total interactions and that bias on the part of these raters affected their estimates of students' writing abilities. Kondo-Brown (2002) examined whether trained native Japanese-speaking raters would rate certain kinds of students and certain criteria more severely or leniently when assessing Japanese L2 writing compositions. She found no systematic overall bias pattern among the raters, but the percentage of significant rater-candidate bias interaction was noticeably higher for candidates of extreme high or low ability. Chiang (1999) studied the importance of various grammatical and discourse features in the evaluation of L2 writing samples by college students enrolled in French classes. The study determined that despite their best efforts, even experienced raters exhibited inconsistency in their

scoring behavior.

This review draws on empirical studies and theoretical discussions from peer-reviewed journals, books, and dissertations published between 1981 and 2024. Databases such as ERIC, JSTOR, and Scopus were searched using terms like "rater bias," "writing assessment," and "EFL." This review examines five broad strategies for addressing rater bias and improving consistency: (1) rater training, (2) use of multiple raters, (3) avoiding rater fatigue, (4) rubric development, and (5) Multi-Faceted Rasch Measurement (MFRM). The central question guiding the review is: What methods have been shown to improve rater consistency in EFL writing assessment across classroom, institutional, and large-scale testing contexts?

1.2 Methodology

A literature search was conducted in ERIC, JSTOR, and Scopus for publications from 1981 through 2024. The primary search strings were "rater bias" AND "EFL writing," "rater consistency" AND "writing assessment," "rater training" AND "EFL writing," "multiple raters" AND "writing assessment," "rater fatigue" AND "writing assessment," "rubric" AND "EFL writing assessment," and "Multi-Faceted Rasch Measurement" AND "rater bias." Studies were included if they examined rater bias, rater consistency, reliability, or strategies for improving the assessment of EFL or L2 writing and were published in peer-reviewed journals, scholarly books, dissertations, or academic conference proceedings. Sources were excluded when they focused exclusively on non-writing assessments, did not address rater-related issues, or lacked sufficient relevance to EFL/L2 writing assessment. Approximately 31 sources were reviewed and synthesized thematically according to five strategies: rater training, multiple raters, reducing rater fatigue, rubric development, and Multi-Faceted Rasch Measurement.

2. Methods Shown to Improve Rater Consistency

Empirical studies have found several ways to improve rater consistency in EFL writing assessment. Rater training, using multiple raters, avoiding rater fatigue, and rubric development have been shown to improve rater consistency (Weigle, 1998; McNamara, 1996; Reddy & Andrade, 2010). It is also recommended that raters and institutions utilize Multi-Faceted Rasch Measurement as a way of detecting sources of rater bias 2008 (Schaefer).

2.1 Rater Training

Rater training refers to structured activities designed to help raters interpret rubrics, apply scoring criteria consistently, and reflect on their own tendencies. Training often includes practice scoring sessions, feedback from expert raters, calibration meetings, and discussion of sample essays at various proficiency levels (Weigle, 1998). The goal is to reduce both intra-rater drift (inconsistency over time) and inter-rater disagreement.

Research has demonstrated that rater consistency can be increased through training (Weigle, 1998). Fulcher and Davidson (2007) stated that the most discussed area of rater training is "the

training of interlocutors for speaking tests and raters for speaking and writing tests" (p. 131). Weigle (2007) admitted that fairly scoring student writing requires raters to have sufficient training and is more difficult than assessing indirect instruments such as grammar and vocabulary tests. She stated that raters trained in EFL writing assessment must be trained to use scoring scales and to apply standards in a consistent, fair manner. Weigle (1999) studied how experienced and inexperienced raters scored essays written by ESL students on two different prompts.

Her analysis showed that inexperienced raters scored more severely on one prompt, but not on the other prompt. After rater training, analysis showed that the differences between the two groups of raters were eliminated following the training. Kondo-Brown (2002) held a training session between ratings on essays with different prompts in her study. Her results found that following her training session, rater's internal consistency improved. Such results demonstrate convincing evidence of the benefits of rater training and its positive impact on rater consistency. These studies suggest that targeted training helps raters interpret rubrics more consistently and mitigates idiosyncratic tendencies.

Other results contradict Kondo-Brown's (2002) study. McNamara (1996) found that rater variability persists despite extensive rater screening and training. Lumley (2002) determined that a tension exists between a rater's internal judgements about a paper and their efforts to apply the grading scale. Lumley (2002) also found that this tension persists despite rater training. Such results serve as a reminder to instructors and institutions alike that while rater training can improve rater consistency, it is nearly impossible to achieve complete rater consistency. Thus, while training is essential, it should be viewed as a necessary but not sufficient step toward consistency.

2.2 Using Multiple Raters

A common solution to variability is to use multiple raters. McNamara (1996) argued that it is unnecessary for a single rater to achieve perfect consistency, since assessment decisions can be strengthened by combining the judgments of multiple raters. When two or more raters score the same essays independently and their scores are averaged or adjudicated, inter-rater reliability improves. This approach does not eliminate individual inconsistency, but it reduces the impact of any single rater's bias.

Other empirical studies support McNamara's (1996) approach. In their study, He et al. (2013) found that raters' academic training and perceptions of category importance were associated with bias patterns. Such findings justify using multiple raters, as each rater carries their own bias when assessing EFL writing, reinforcing the value of multiple perspectives. Kim et al. (2017) examined how raters and tasks influence measurement error in writing evaluation in Grade 3 and Grade 4 students. Also examined was how many raters and tasks were needed to reach a desirable and strong reliability score of .90 and .80. Kim et al. (2017) found that to reach the reliability of .90, multiple tasks and raters were needed indicating that using multiple raters does

have a positive effect on the reliability of writing assessment. Institutions should utilize multiple raters when assessing an EFL writing piece in order to ensure rater consistency. In institutional EFL writing assessment, two or more raters should independently assign scores to assure reliability, especially if this assessment takes place on a department-wide or institution-wide basis (Ferris & Hedgcock, 2014, p. 208).

2.3 Avoiding Rater Fatigue

While multiple raters can help improve reliability, fatigue remains an important threat to consistency in writing assessment. Rater fatigue is not only about being physically tired but also about the mental strain of scoring long sessions of essays. Mahshanian and Shahnazari (2020) showed that comment frequency declined over time, indicating that raters tended to give richer feedback at the beginning of a session but less toward the end. This pattern highlights how fatigue can subtly shape scoring behavior in ways that undermine fairness.

Wind and Schumacker (2017) expanded on this idea by framing fatigue as a form of "measurement disturbance." They argued that disturbances such as boredom, plodding, or overexertion can cause raters to deviate from expected scoring patterns. Importantly, their study demonstrated that these problems often remain invisible when using only conventional fit statistics. By applying graphical Rasch methods, they showed that it is possible to detect hidden disturbances that signal fatigue or inconsistency. This makes a strong case for institutions to adopt more sophisticated monitoring systems to ensure that raters maintain consistent judgment across sessions.

Task design also plays a role in fatigue. Naismith, Attali, and LaFlair (2024) observed that longer writing tasks may increase the cognitive demands on raters, raising the likelihood of inconsistency. They investigated this question more directly by comparing short (5-minute) and longer (20-minute) writing tasks. Their results revealed that longer durations did not improve reliability or validity but did yield lengthier essays that were more taxing to evaluate. This finding implies that assessment developers can shorten tasks without sacrificing measurement quality, thereby reducing rater fatigue and allowing more efficient scoring.

Taken together, the evidence suggests that avoiding rater fatigue requires both procedural and design-based strategies. Institutions should consider limiting the number of scripts scored in a single sitting, structuring breaks into scoring sessions, and monitoring rater performance for hidden disturbances. At the same time, test developers should be cautious about assigning overly long tasks, since they may add cognitive burden without improving validity. By addressing both the human and design factors that contribute to fatigue, assessment programs can take meaningful steps toward reducing bias and strengthening the fairness of EFL writing evaluation.

2.4 Rubric Development

"Fair, informative writing assessment practices take into account students' needs and competencies, social and institutional expectations regarding writing performance, target genres,

and the teacher's readiness to deploy assessment tools, and the quality of the assessment instruments themselves" (Ferris & Hedgcock, 2014, p. 203). A well-constructed rubric is central to reliable writing assessment. Rubrics clarify expectations, promote fairness, and provide a common framework for raters and students. However, rubrics vary widely in quality, and poorly designed rubrics can exacerbate inconsistency rather than reduce it. Analytic rubrics, which break down writing into components such as grammar, vocabulary, organization, and content, are generally considered more reliable than holistic rubrics, which assign one overall score.

A critical review of empirical research surrounding the use of rubrics at the post-secondary level found that studies of rater reliability tended to show rubrics can lead to a relatively common interpretation of student performance (Reddy & Andrade, 2010). Scoring rubrics have traditionally been used to aid raters, but they can also help students and should be used systematically during the assessment process (Ferris & Hedgcock, 2014, p. 207). Park (2012) examined the differences between native and non-native raters when assessing EFL writing. She found and stated in her conclusion that developing an appropriate rubric has the potential to make the raters on writing performance have greater inter-rater reliability. Li (2018) studied the use of rubrics in university-level EFL writing classes in China and examined the impact of teacher training on the application of rubrics. The implementation study showed an average positive increase in the writing performance of Chinese EFL college students after rubrics were used as an assessment tool. Additionally, the raters in Li's (2018) study held positive attitudes towards rubrics as they were able to improve the consistency of their ratings. Hamp-Lyons (1991) argued that any method of writing assessment that "fails to utilize the educative potential" of the instrument itself "permits a disjunction between teaching and assessment, a disjunction we have suffered under for all too long and need suffer no longer" (p. 244).

However, rubrics also present challenges. While rubrics can improve rater consistency, they cannot eliminate bias altogether. For example, Hamp-Lyons (1996) warned that analytic rubrics may bias raters toward surface features that are easy to identify, rather than deeper qualities like argumentation. Crusan (2010) argued that rubrics should be developed to reflect classroom and institutional goals, stressing that "all assessment should be locally designed and locally controlled," though at the same time admitted that constructing rubrics and scales is "one of the most worrisome and … overwhelming" assessment responsibilities (p. 72). Ferris and Hedgcock (2014) agreed with Crusan's (2010) argument and recommended a balanced approach: adapting established rubrics while tailoring them to local contexts. Examples of effective rubrics include those with clear, parallel descriptors across proficiency levels, while ineffective rubrics may contain vague, overlapping, or culturally biased language. They advised that instructors should consider "online rubric-generating tools such as Rubistar and TeacherVision" (p. 215).

2.5 Multi-Faceted Rasch Measurement

Multi-Faceted Rasch Measurement (MFRM) is a statistical model that extends Rasch measurement to analyze multiple sources of variability simultaneously including raters, tasks,

and test-takers. Instead of assuming that test scores reflect only student ability, MFRM isolates the contributions of each "facet" to the score, allowing researchers and institutions to identify sources of bias (Myford & Wolfe, 2004).

Multiple studies have proven the efficacy of multi-faceted Rasch measurement (MFRM) as a way of analyzing sources of variability in writing test scores due to rater biases (He et al., 2013; McNamara, 1996; Schaefer, 2008). Schaefer (2008) employed MFRM to investigate the rater bias patterns of native English-speaking raters when they rate EFL essays. His results revealed several reoccurring bias patterns among the raters in his study. For example, he found that in rater-writer bias interactions, there tended to be more severe or lenient bias towards higher ability writers than writers with lower ability. Farrokhi et al. (2012) was able to use MFRM to investigate rater severity or leniency among three rater types including self-assessor, peer-assessor, and teacher assessor. MFRM was able to reveal differing patterns of severity among the three assessment types. For example, the study found that those in the self-assessor and teacher assessor categories demonstrated the opposite pattern of severity and leniency compared to the peer-assessor group when assessing the highest and lowest ability students.

He et al. (2013) applied MFRM to examine rater bias in the assessment of essays written by EFL college students. They found that argumentative essays were rated more severely than descriptive essays. In addition, He et al. (2013) used MFRM to determine that raters' academic training and perceptions of category importance were associated with their bias patterns.

The advantage of MFRM is that it provides actionable data: institutions can identify raters whose scoring patterns deviate significantly and provide targeted retraining or adjust scores statistically. However, MFRM has limitations. It requires specialized knowledge and software, and it can detect bias but cannot prevent it. Some question whether additional measures to improve rater consistency are necessary if MFRM already identifies rater bias. Supporters argue that MFRM can address rater inconsistency. The method assumes that raters judge independently but from the "same point of view" (Linacre, 1998, p. 630). MFRM reveals whether this condition holds and identifies rater misfit. In classroom settings, misfitting raters can invalidate test results, particularly when only a few raters are used. While MFRM does not solve this problem directly, it provides information that can guide targeted feedback, improve rater training, and enhance rater consistency (Myford & Wolfe, 2004, p. 564).

3. Key Findings

Table 1 summarizes the key findings, limitations, and contributions of the major studies reviewed, organized by the method they investigated for improving rater consistency.

Table 1. Comparative table

Method	Study	Main Findings	Limitations/Critiques
Rater Training	Weigle (1998, 1999, 2002, 2007)	Training (practice, calibration) increases consistency; inexperienced raters converge after training.	Training helps but does not eliminate variability; resource-intensive.
	Kondo-Brown (2002)	Post-training internal consistency improved among raters.	Small sample; did not remove all bias patterns.
	Fulcher & Davidson (2007)	Emphasize structured training and calibration for speaking/writing tests.	Implementation varies; practical constraints.
	Lumley (2002)	Raters experience tension between rubric rules and internal judgment; training doesn't fully remove this.	Suggests limits to training.
	Chiang (1999)	Even experienced raters show inconsistency judging grammatical/textual features.	Context-specific (French L2 classroom).
Multiple Raters	McNamara (1996)	Combining multiple raters reduces influence of individual bias.	Adds logistical/financial burden; doesn't remove all systematic sources.
	Du, Wright & Brown (1996)	Detected differential-facet effects (rater × candidate interactions).	Context-specific; suggests need for monitoring.
	He et al. (2013)	Bias patterns tied to training/topic; multiple raters help mitigate.	Some bias remains even with many raters.
	Kim et al. (2017)	Multiple raters/tasks needed for high reliability (.90/.80).	Requires substantial resources.
	Ferris & Hedgcock (2014)	Recommend 2+ raters in institutional writing assessment to boost fairness.	Feasibility depends on institutional capacity.
Avoiding Rater Fatigue	Mahshanian & Shahnazari (2020)	Rater fatigue reduces comment frequency and alters scoring patterns across a session.	Small sample/context limits generalizability.
	Wind & Schumacker (2017)	Identified rater fatigue, boredom, and plodding as measurement disturbances; graphical Rasch displays detect disturbances missed by fit statistics.	Requires technical expertise; not widely used.
	Naismith, Attali & LaFlair (2024)	Longer tasks increase rater cognitive load without improving reliability; shorter tasks maintain score quality and reduce fatigue.	Evidence limited to short vs. long tasks in one testing context.
Rubric Development	Ferris & Hedgcock (2014)	Rubrics clarify expectations, support rater agreement, and help students.	Effectiveness depends on rubric design/implementation.
	Reddy & Andrade (2010)	Meta-review: rubrics generally improve rater consistency and clarity.	Findings heterogeneous across studies.
	Li (2018)	Locally-developed rubrics improved student performance and rater attitudes.	Small-scale, local context.

Method	Study	Main Findings	Limitations/Critiques
	Sundeen (2014)	Rubric presentation/design affects rater use and writing quality.	Effects depend on context and presentation options.
	Zhang (2012)	Rubric assessment supports clearer expectations in English writing courses.	Conference paper; limited empirical breadth.
	Hamp-Lyons (1991, 1996)	Analytic rubrics increase transparency but may bias raters toward surface features.	Risk of over-emphasizing easily measurable elements.
	Crusan (2010)	Advocates for locally tailored rubrics and thoughtful design.	Time-consuming to design/validate; limited generalizability.
Multi-Faceted Rasch Measurement (MFRM)	Schaefer (2008)	MFRM identifies recurring rater bias patterns (severity/leniency).	Identifies bias but doesn't remove it; needs expertise.
	Farrokhi, Esfandiari & Shaefer (2012)	MFRM differentiates rater severity across assessor types (teacher/peer/self).	Sample-specific results; technical interpretation required.
	Myford & Wolfe (2004)	Practical guidance for detecting and measuring rater effects with MFRM.	Useful diagnostically but not a standalone fix for inconsistency.
	Linacre (1998)	Theoretical notes on rating, judges, fairness and Rasch considerations.	Theoretical framing rather than empirical test of solutions.
	Park (2012)	MFRM applied to EFL writing to detect rater bias in a local context.	Context-specific; supports MFRM's local utility.
	He et al. (2013)	(Also under Multiple Raters) MFRM shows topic & training influence bias.	Requires many-facet data and expertise.
Other/ Foundational	Shepard, Camilli & Averill (1981)	Early comparison of methods for detecting item/test bias — foundational to bias detection.	Older methods; still conceptually relevant.
	McNamara (1996)	Classic overview of measuring L2 performance and rater issues.	Broad textbook; not a single empirical study.
	Katz (2014)	Perspectives on validity and fairness in classroom assessment.	Conceptual rather than empirical.
	Klenowski (2009)	Discusses assessment for learning and teacher judgment.	Theoretical framing, not rater-specific.

4. Discussion

4.1 Rater Bias Across Assessment Contexts

Rater bias appears differently depending on the context in which EFL writing is assessed. Although the underlying problem is variation in rater judgment, the scale and purpose of assessment influence how bias manifests and which strategies are practical for addressing it. The

literature reviewed in this paper suggests important distinctions among classroom, institutional, and large-scale assessment contexts.

In classroom assessment, bias is often closely connected to individual teacher judgment, local instructional goals, and the subjective interpretation of writing quality. Zhang (2012) found that college-level EFL writing assessment could be ambiguous and strongly influenced by subjective factors, while Katz (2014) emphasized the relationship between classroom assessment, instructional planning, and feedback. Because classroom assessment is typically conducted by a relatively small number of teachers, individual rater tendencies can have a direct effect on students' scores. Rubrics and rater training are therefore particularly relevant in this context. Locally designed rubrics can clarify expectations and provide teachers with a consistent framework for evaluating writing, although they must remain appropriate to classroom and institutional goals (Crusan, 2010; Ferris & Hedgcock, 2014). Training can similarly help teachers apply scoring criteria more consistently, but research indicates that training does not completely eliminate individual differences in judgment (Lumley, 2002; Weigle, 1998).

In institutional assessment, the potential influence of individual raters becomes more significant because writing may be assessed across courses, programs, or departments. Ferris and Hedgcock (2014) recommend using two or more raters for institutional writing assessment, particularly when assessment occurs on a department-wide or institution-wide basis. Multiple raters can reduce the influence of any one individual's scoring tendencies, while shared rubrics and calibration activities can establish greater consistency across raters. Institutional assessment also creates opportunities for systematic monitoring of rater performance. MFRM can be particularly useful because it allows institutions to identify differences in rater severity or leniency and examine interactions among raters, tasks, and test takers (He et al., 2013; Myford & Wolfe, 2004; Schaefer, 2008). Thus, institutional management of bias can move beyond individual teacher training toward coordinated procedures involving multiple raters, common scoring criteria, and statistical monitoring.

In large-scale assessment, the effects of rater variability become more consequential because writing scores may be used for high-stakes decisions involving large numbers of test takers. McNamara (1996) demonstrated that rater variability can persist even when raters are carefully selected and trained, while Lumley (2002) found that raters may interpret assessment criteria differently despite efforts to standardize scoring. Large-scale assessment therefore requires procedures that can manage variability across many raters and large numbers of writing samples. Multiple raters, standardized scoring criteria, and systematic rater training can reduce the influence of individual judgment. Rater fatigue is also particularly relevant when raters score large numbers of responses, as fatigue can alter scoring behavior and feedback patterns (Mahshanian & Shahnazari, 2020; Wind & Schumacker, 2017). MFRM provides an additional mechanism for detecting systematic differences in rater severity and identifying unusual scoring patterns that may otherwise remain hidden (Schaefer, 2008).

The distinction among these contexts does not mean that particular strategies belong exclusively to one level. Rather, the scale of assessment changes how the strategies are implemented and how important systematic monitoring becomes. In classroom settings, the primary concern may be helping individual teachers recognize and manage their own scoring tendencies through training and appropriate rubric use. At the institutional level, the emphasis shifts toward coordinating multiple raters and establishing shared assessment procedures. In large-scale testing, these practices must be supplemented by more extensive monitoring, fatigue management, and statistical analysis of rater effects. Across all three contexts, the literature indicates that no single strategy can eliminate bias; instead, effective management requires combining procedures appropriate to the scale and purpose of the assessment.

4.2 Intra-Rater and Inter-Rater Consistency

Another important distinction in understanding rater bias is the difference between intra-rater and inter-rater consistency. Intra-rater consistency refers to the extent to which the same rater applies scoring criteria consistently across different writing samples, prompts, or points in time. Inter-rater consistency, in contrast, concerns the extent to which different raters apply the same criteria similarly when evaluating the same writing. The five strategies reviewed in this paper address these two forms of consistency to different degrees. Some methods primarily target individual rater behavior, while others reduce the effects of differences between raters.

Rater training has implications for both intra-rater and inter-rater consistency. Training can help individual raters develop a more stable interpretation of scoring criteria, thereby reducing inconsistency within a rater's own scoring over time. Weigle (1998, 1999) found that training reduced differences between experienced and inexperienced raters and helped address differences associated with writing prompts. Kondo-Brown (2002) similarly found improved internal consistency following a training session. At the same time, calibration and discussion of sample essays can help different raters develop more similar interpretations of a rubric, improving inter-rater consistency (Weigle, 1998; Fulcher & Davidson, 2007). However, the evidence also indicates that training does not eliminate either form of inconsistency. Lumley (2002) found that raters continued to experience tension between their individual judgments and the requirements of the scoring criteria, while McNamara (1996) found that rater variability could persist despite extensive training. Thus, training can support both forms of consistency, but its effects depend on the extent to which raters internalize and apply shared assessment criteria.

Using multiple raters primarily addresses inter-rater consistency rather than intra-rater consistency. When two or more raters independently assess the same writing, differences in individual severity or leniency can be identified, averaged, or adjudicated, reducing the influence of any single rater's judgment (McNamara, 1996). He et al. (2013) found that rater characteristics and perceptions of scoring categories were associated with different bias patterns, demonstrating why reliance on a single rater can be problematic. Kim et al. (2017) likewise found that the number of raters and writing tasks affected the reliability of writing scores.

Multiple raters, however, do not necessarily make each individual rater more internally consistent. Instead, they provide a mechanism for reducing the effect of individual differences between raters. For this reason, multiple-rater procedures are particularly important when assessment decisions depend on agreement across raters or when writing is assessed at an institutional or large-scale level.

Avoiding rater fatigue is particularly relevant to intra-rater consistency because fatigue can cause an individual rater's scoring behavior to change during an assessment session. Mahshanian and Shahnazari (2020) found that comment frequency declined over the course of scoring, suggesting that rater behavior may change as fatigue increases. Wind and Schumacker (2017) similarly identified fatigue, boredom, and overexertion as potential measurement disturbances. These effects can also have consequences for inter-rater consistency if different raters experience fatigue at different points or to different degrees. However, the most immediate concern is maintaining stable judgment within an individual rater across a scoring session. Limiting the number of writing samples scored at one time, incorporating breaks, and reducing unnecessary task demands can therefore help preserve intra-rater consistency (Mahshanian & Shahnazari, 2020; Naismith et al., 2024).

Rubric development can support both intra-rater and inter-rater consistency by providing a stable framework for evaluating writing. For individual raters, clearly defined criteria can reduce uncertainty and help maintain consistent judgments across writing samples. For multiple raters, shared criteria can provide a common basis for scoring and reduce differences in how raters interpret the construct being assessed. Reddy and Andrade (2010) found that rubrics can promote a relatively common interpretation of student performance, while Park (2012) suggested that appropriate rubric development can increase inter-rater reliability. Nevertheless, rubrics do not guarantee either form of consistency. Hamp-Lyons (1996) cautioned that analytic rubrics may encourage raters to focus on surface features that are easier to identify, while Crusan (2010) emphasized the importance of designing assessment criteria that reflect local instructional and institutional purposes. Consequently, rubric quality and appropriate implementation are important for maintaining consistency within and across raters.

Multi-Faceted Rasch Measurement (MFRM) differs from the other methods because it is primarily a diagnostic rather than a preventive approach. MFRM can identify patterns of rater severity and leniency and examine interactions among raters, writing tasks, and test takers (Myford & Wolfe, 2004). This makes it useful for investigating inter-rater differences, including whether particular raters consistently score more severely or leniently than others. At the same time, MFRM can identify patterns within an individual rater's scoring behavior, allowing institutions to investigate whether a rater's judgments vary systematically across different types or levels of writing. Schaefer (2008), for example, used MFRM to identify recurring rater bias patterns, while He et al. (2013) found that topic and rater characteristics were associated with differences in scoring. MFRM therefore provides information relevant to both intra-rater and inter-rater consistency, although it does not itself eliminate the source of the inconsistency.

Instead, its results can inform targeted training, monitoring, and other corrective procedures.

Taken together, the five methods address intra-rater and inter-rater consistency through different mechanisms. Rater training and rubric development can contribute to both forms of consistency, while multiple raters primarily reduce the consequences of inter-rater differences and fatigue management primarily protects consistency within individual raters. MFRM provides a means of identifying patterns in both types of inconsistency. The distinction is therefore important because improving agreement between raters does not necessarily mean that each rater is internally consistent, just as improving an individual rater's consistency does not ensure that different raters will reach similar judgments. A comprehensive approach to EFL writing assessment should consequently consider both forms of consistency when selecting and combining strategies.

4.3 Implications for Improving Rater Consistency

An analysis of the literature reveals implications for classroom assessment and the field of assessment research more broadly. The most notable finding is that one method alone has not been proven to achieve complete rater consistency. Advocates of specific methods such as rater training and using multiple raters debate the extent to which a method improves rater consistency, however, all admit that it is not possible to reach complete consistency when assessing EFL writing. As a result, it is not advisable for an instructor or institution to only utilize one of the methods reviewed in this review. Rather, it is recommended that instructors and institutions utilize a combination of the practices discussed in this paper in order to improve rater consistency. For example, a combination of rater training and utilizing multiple raters can increase rater consistency when assessing EFL writing. Individual rater training will improve internal rater consistency while utilizing multiple raters will mitigate the bias patterns of each rater (He et al., 2013; Kondo-Brown, 2002).

While the extent to which rater training, avoiding rater fatigue, using multiple raters, and rubric development reduce rater inconsistency remain subject to much debate, the evidence surrounding the efficacy of multi-faceted Rasch measurement (MFRM) as a way of analyzing sources of variability in writing test scores due to rater biases remains consistent (He et al., 2013; McNamara, 1996; Schaefer, 2008). It is important for raters and institutions to remember that while MFRM expects raters to assess a written piece independently, it assumes that they will still judge from "the same point of view." Raters and institutions should also remember that MFRM cannot solve the problem of bias in EFL writing assessment. Rather it provides information that can be used for targeted feedback, better rater training, and improve rater consistency (Myford & Wolfe, 2004, p. 564).

The contribution of this review lies in its synthesis of five approaches to improving rater consistency within a comparative framework. Rather than examining rater training, multiple raters, fatigue management, rubric development, and Multi-Faceted Rasch Measurement (MFRM) as isolated strategies, this review considers how each approach addresses different sources and forms of rater inconsistency across assessment contexts. Table 1 provides a comparative

synthesis of the reviewed studies by identifying their primary findings, practical applications, and limitations. Together, this framework and comparative analysis show that the five approaches serve different functions and can be used in combination rather than as competing solutions. This synthesis provides a more integrated understanding of how strategies for improving rater consistency can be selected and combined according to the assessment context and the type of inconsistency being addressed.

Few studies directly compare combinations of methods such as rubric use and MFRM. Moreover, the role of cultural background in rater bias remains underexplored. Researchers should examine which combinations of the methods discussed are most effective at improving rater consistency. For example, the relationship between rater training and rubric development has implications for classroom assessment and institutions. Future studies could examine how rubric design influences raters of different cultural backgrounds, or whether training programs tied to specific rubrics outperform generalized approaches. Some institutions may train raters to assess writing with a specific rubric, while others may have more generalized training programs. A potential topic for investigation is how the type of training program affects rater consistency. Researchers should also explore how multi-faceted Rasch measurement (MFRM) can be used to reduce rater inconsistency in EFL writing assessment. Given the multifaceted nature of rater inconsistency, no single intervention is sufficient. Institutions should adopt layered approaches—such as pairing training with statistical monitoring tools like MFRM—to address both overt and subtle forms of bias.

References

Chiang, S. Y. (1999). Assessing grammatical and textual features in L2 writing samples: The case of French as a foreign language. *The Modern Language Journal, 83*(2), 219-232. https://doi.org/10.1111/0026-7902.00017

Crusan, D. (2010). *Assessment in the second language writing classroom.* University of Michigan Press. https://doi.org/10.3998/mpub.770334

Du, Y., Wright, B. D., & Brown, W. L. (1996). *Differential facet functioning detection in direct writing assessment [Paper presentation].* Annual Meeting of the American Educational Research Association, New York, NY, United States.

Farrokhi, F., Esfandiari, R., & Schaefer, E. (2012). A many-facet Rasch measurement of differential rater severity/leniency in three types of assessment. *JALT Journal, 34*(1), 79-101. https://doi.org/10.37546/JALTJJ34.1-3

Ferris, D. R., & Hedgcock, J. S. (2014). *Teaching ESL composition: Purpose, process, and practice* (3rd ed.). Routledge. https://doi.org/10.4324/9781410611505

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book.* Routledge. https://doi.org/10.4324/9780203449066

Hamp-Lyons, L. (1991). Scoring procedures for ESL contexts. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 241-276). Ablex.

Hamp-Lyons, L. (1996). The challenges of second-language writing assessment. In E. M. White, W. D. Lutz, & S. Kamusikiri (Eds.), *Assessment of writing: Politics, policies, and practices* (pp. 226-240). Modern Language Association.

He, T.-H., Gou, W. J., Chien, Y.-C., Chen, I.-S. J., & Chang, S.-M. (2013). Multi-faceted Rasch measurement and bias patterns in EFL writing performance assessment. *Psychological Reports, 112*(2), 469-485. https://doi.org/10.2466/03.11.PR0.112.2.469-485

Katz, A. (2014). Assessment in second language classrooms. In M. Celce-Murcia, D. M. Brinton, & M. A. Snow (Eds.), *Teaching English as a second or foreign language* (4th ed., pp. 320-337). National Geographic Learning.

Kim, Y.-S. G., Schatschneider, C., Wanzek, J., Gatlin, B., & Al Otaiba, S. (2017). Writing evaluation: Rater and task effects on the reliability of writing scores for children in grades 3 and 4. *Reading and Writing, 30*(6), 1287-1310. https://doi.org/10.1007/s11145-017-9724-6

Klenowski, V. (2009). Assessment for learning revisited: An Asia-Pacific perspective. *Assessment in Education: Principles, Policy & Practice, 16*(3), 263-268. https://doi.org/10.1080/09695940903319646

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese L2 writing performance. *Language Testing, 19*(1), 3-31. https://doi.org/10.1191/0265532202lt218oa

Li, C. (2018). *Constructing and applying rubrics in college-level EFL writing assessment in China* (Master's thesis, Syracuse University, Syracuse, New York, United States). Retrieved from https://surface.syr.edu/thesis/241

Linacre, J. M. (1998). Rating, judges, and fairness. *Rasch Measurement Transactions, 12*(2), 630-631. Retrieved from http://www.rasch.org/rmt/rmt122f.htm

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing, 19*(3), 246-276. https://doi.org/10.1191/0265532202lt230oa

Mahshanian, A., & Shahnazari, M. (2020). The effect of raters fatigue on scoring EFL writing tasks. *Indonesian Journal of Applied Linguistics, 10*(1), 1-13. https://doi.org/10.17509/ijal.v10i1.24956

McNamara, T. F. (1996). *Measuring second language performance.* Longman.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. In E. V. Smith, Jr., & R. M. Smith (Eds.), *Introduction to Rasch measurement: Theory, models and applications* (pp. 518-574). JAM Press.

Naismith, B., Attali, Y., & LaFlair, G. T. (2024). The impact of task duration on the scoring of

independent writing responses of adult L2-English writers. *Assessing Writing, 62*, 100895. https://doi.org/10.1016/j.asw.2024.100895

Park, M. (2012). Exploring the raters' bias on an EFL writing assessment using multi-faceted Rasch measurement. *Studies in English Education, 17*(2), 175-202.

Reddy, Y. M., & Andrade, H. (2010). A review of rubric use in higher education. *Assessment & Evaluation in Higher Education, 35*(4), 435-448. https://doi.org/10.1080/02602930902862859

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing, 25*(4), 465-493. https://doi.org/10.1177/0265532208094273

Shepard, L., Camilli, G., & Averill, M. (1981). Comparison of procedures for detecting test-item bias with both internal and external ability criteria. *Journal of Educational and Behavioral Statistics, 6*(4), 317-375. https://doi.org/10.3102/10769986006004317

Sundeen, T. (2014). Instructional rubrics: Effects of presentation options on writing quality. *Assessing Writing, 21*, 74-88. https://doi.org/10.1016/j.asw.2014.03.003

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

Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing, 6*(2), 145-178. https://doi.org/10.1016/S1075-2935(00)00010-6

Weigle, S. C. (2002). *Assessing writing.* Cambridge University Press. https://doi.org/10.1017/CBO9780511732997

Weigle, S. C. (2007). Teaching writing teachers about assessment. *Journal of Second Language Writing, 16*(3), 194-209. https://doi.org/10.1016/j.jslw.2007.07.004

Wind, S. A., & Schumacker, R. E. (2017). Detecting measurement disturbances in rater-mediated assessments. *Educational Measurement: Issues and Practice, 36*(4), 44-51. https://doi.org/10.1111/emip.12164

Zhang, L. (2012). Rubric assessment in English writing courses. *Proceedings of the Second Northeast Asia International Symposium on Language, Literature, and Translation, 6*, 217-222.

Author Statements

Acknowledgments

Not applicable.

Author Contributions

Not applicable.

Consent for Publication

All authors have read and approved the published version of the manuscript.

Funding

This research received no external funding.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Informed Consent

Not applicable.

Ethics Approval

Ethics approval was not required because manuscript serves as a meta-anaylsis.

Data Availability Statement

Not applicable: No new data were created or analyzed in this study.

Use of Artificial Intelligence

The authors declare that no generative artificial intelligence (AI) tools were used in the conception, design, analysis, interpretation, or writing of this manuscript.

Supplementary Materials

No supplementary materials are available.

Journal Information

The Editorial Board declared the below statement on behalf of the journal and publisher.

Ethics Statement

The journal follows publication-ethics policies informed by current COPE guidance.

Preservation Policy

This journal is archived by LOCKSS Archiving.

Open Access

This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0). Please find details at https://creativecommons.org/licenses/by/4.0.

Copyright © The Author(s)

Copyright remains with the author(s). The author(s) grant the journal the right of first publication, and the article is published under the terms of the Creative Commons Attribution (CC BY 4.0) License.