**The Journal of English Language Pedagogy and Practice**
Vol. 17, No.35, Autumn and Winter 2024
DOI: 10.71586/jal.2024.11181190893

Research Article

# Developing a Diagnostic-Oriented Scale for EFL Academic Writing: An Empirical Approach

Fatemeh Shoaei[1], Sayyed Mohammad Alavi[2*], Hossein Karami[3]
[1]*Department of English Language and Literature, Alborz Campus, University of Tehran, Tehran, Iran*
[2,3]*Faculty of Foreign Languages and Literatures, University of Tehran, Tehran, Iran*

*Corresponding author: smalavi@ut.ac.ir

(Received: 2024/11/18; Accepted: 2025/02/22)
Online publication: 2025/02/27

## Abstract

Despite the growing interest in diagnostic assessment tools in second language writing, empirical research on their development for EFL contexts remains limited. This study responds to this gap by developing and validating a diagnostic-oriented rating scale tailored for Iranian EFL learners' academic writing. Employing a mixed-methods approach, essential descriptors reflecting core writing skills were identified through thematic analysis of the data gathered from think-aloud protocols and expert feedback. These descriptors underwent rigorous statistical validation, including Intraclass Correlation Coefficient (ICC) for inter-rater reliability, Content Validity Index (CVI) for content validity, and Pearson correlation coefficients for concurrent validity. The findings indicate that the 21 empirically derived descriptors effectively capture crucial aspects of academic writing—content fulfillment, organizational knowledge, and language usage—thereby enabling instructors to assess learner proficiency with greater precision. The scale provides substantial value for both large-scale assessments and classroom applications, fostering a learner-centered approach that empowers students to identify and overcome specific writing challenges.

***Keywords:*** scale development, diagnostic assessment, EFL academic writing, empirical approach, descriptor

**The Journal of English Language Pedagogy and Practice**
Vol. 17, No.35, Autumn and Winter 2024
DOI: 10.71586/jal.2024.11181190893

## Introduction

The assessment of academic writing skills in EFL (English as a Foreign Language) learners has become increasingly important, particularly due to the global demand for English proficiency. Effective writing assessment is crucial not only for academic success but also for equipping EFL learners with the skills necessary for higher education and professional contexts. English proficiency, especially in writing, is often seen as a gateway to broader academic and professional opportunities worldwide (Brown & Harris, 2016; Kellogg & Raulerson, 2007). However, traditional assessment methods often fall short in providing fine-grained, diagnostic feedback that can guide learners in identifying specific areas of strength and improvement (Safari & Ahmadi, 2023; Weigle, 2002). While these methods may assess overall competence, they do not offer the targeted, detailed feedback that can help learners address specific weaknesses. This limitation has led to the need for more precise and nuanced diagnostic tools that can accurately capture the complexities of academic writing and offer targeted feedback for learners and instructors alike.

Despite the recognized importance of diagnostic assessment, there remains a dearth of empirical research focused on developing diagnostic scales specifically tailored for EFL writing contexts. Although recent studies (e.g., He, Jiang, & Min, 2021; Khamboonruang, 2022; Ma, Shi, Lu, & Li, 2022; Safari & Ahmadi, 2023) have made notable contributions, the field still lacks a diverse range of empirically validated, context-sensitive diagnostic tools capable of ensuring consistency and accuracy in assessment. This gap in EFL-specific tools limits the ability of educators to provide targeted feedback that aligns with EFL learners' unique challenges, ultimately restricting both learners and instructors from accessing consistent, reliable assessments of academic writing proficiency. Research by Alderson, Brunfaut, and Harding (2015) and Harding, Alderson, and Brunfaut (2015) underscore the role of diagnostic tools in providing feedback that supports learners' skill development over time. Although diagnostic assessment is valued for helping learners identify and address specific weaknesses, many existing scales lack a comprehensive empirical foundation and validation process, which can result in limited reliability in assessment outcomes. Foundational studies, including those by Fulcher (1993), Upshur and Turner (1995), North (2003), Knoch (2009), and more recent research by Kim (2019), He et al. (2021), and Safari and Ahmadi (2023), emphasize the need for rating scales to be grounded in empirical data to improve their reliability and validity. Kim (2019) also critiques

**The Journal of English Language Pedagogy and Practice**
Vol. 17, No.35, Autumn and Winter 2024
DOI: 10.71586/jal.2024.11181190893

the reliance on intuitive or a priori approaches, noting that scales lacking empirical validation can undermine assessment consistency. This gap in research highlights the need for empirically-derived diagnostic tools tailored to the EFL writing context, which could facilitate more nuanced and effective feedback.

Traditional approaches to writing assessment in EFL contexts, such as holistic and analytic rubrics, have long been used due to their simplicity and ability to provide quick evaluations. However, these methods are often criticized for their limitations in diagnostic accuracy. An evolving debate (e.g., Hamp-Lyons, 1995, 2016; Park & Yan, 2019; Weigle, 2002; Zou, Yan, & Fan, 2024) has examined how holistic scoring, which provides a single score based on an overall impression, tends to obscure specific strengths and weaknesses. This approach leaves learners with little understanding of where they need to improve. Such broad scoring methods can lead to subjective interpretations, resulting in inconsistent feedback across diverse writing samples, as they lack precision in evaluating specific aspects of writing performance.

Even with more detailed analytic rubrics, where individual components of writing (such as grammar, coherence, and vocabulary) are separately scored, challenges persist. Perkins (1983) has noted that these scales tend to evaluate broad categories, offering general subscores that may obscure specific writing issues, as they typically isolate features from context and lack flexibility to account for variations in discourse types. This drawback can hinder their effectiveness in providing targeted feedback for nuanced improvements.

These limitations highlight the need for more sophisticated diagnostic tools that can offer detailed, meaningful feedback, allowing learners to identify and improve specific skills. Rupp, Templin, and Henson (2010) argue for a shift away from traditional assessment methods by advocating for diagnostic approaches that provide a more precise understanding of individual skill areas. Such tools are essential for supporting targeted interventions and personalized learning, as they offer richer, more specific insights into learners' strengths and weaknesses, enabling educators to tailor instruction to address specific learning needs effectively.

## Diagnostic Assessment in Writing

Over the past decade, diagnostic assessment has garnered increased attention from educational experts as an alternative to traditional feedback methods because of its capacity to deliver detailed insights into students' strengths and weaknesses (Alderson et al., 2015). It involves the use of assessments specifically designed to identify areas where learners excel or struggle, thereby guiding decisions on future

**The Journal of English Language Pedagogy and Practice**
Vol. 17, No.35, Autumn and Winter 2024
DOI: 10.71586/jal.2024.11181190893

teaching, training, or learning (Lu, Han, Fang, & Shen, 2021). This approach effectively integrates assessment with instruction, making it particularly useful for pinpointing student challenges and offering targeted solutions (Rupp et al., 2010). While beneficial for all students, diagnostic feedback and correction are particularly valuable for second language writers (Mäkinen, 1995).

In the last few years, research on diagnostic writing assessment has outlined essential principles, systematic procedures, and a range of effective tools designed to produce a detailed profile of language learners' issues and challenges (e.g., He et al., 2021; Khamboonruang, 2020; Lu et al., 2021; Safari & Ahmadi, 2023; Wang & Xie, 2022). For instance, Khamboonruang (2020) developed and validated a diagnostic rating scale for formative assessment in EFL university writing classrooms. Constructed through a multi-stage exploratory mixed-methods approach, the scale drew on L2 writing theories, existing scales, expert input, and classroom curricula. Over the course of a semester, the scale was implemented with 80 undergraduate students and five instructors, enabling the diagnosis of students' writing strengths and weaknesses and providing targeted feedback to support ongoing teaching and learning. While the study provided reasonable support for the diagnostic system, further evidence was needed to reliably track long-term student progress across different writing tasks. These findings underscored the need for continuous validation and refinement of diagnostic tools to address the complexities of EFL writing development.

Focusing on integrated writing tasks, Safari and Ahmadi (2023) developed and validated an empirically-based binary-choice diagnostic checklist designed to assess L2 students' performance in reading-listening-writing tasks. The checklist consisted of 30 one-sentence descriptors targeting specific aspects of integrated essays, which were revised by two ESL writing experts. Their study demonstrated that the checklist provided detailed diagnostic feedback, helping to pinpoint students' strengths and weaknesses across various writing components. The findings revealed high score consistency among raters across multiple prompts, with raters confidently applying the checklist without difficulty. This diagnostic tool expanded the literature by highlighting the benefits of descriptor-based, binary-choice diagnostic checklists for integrated writing assessment.

## Development of Writing Scales: Approaches and Challenges

The development of diagnostic writing scales involves several complex challenges, ranging from ensuring precise, analytical descriptors to balancing comprehensive feedback with practical usability. Research suggests that instead

**The Journal of English Language Pedagogy and Practice**
Vol. 17, No.35, Autumn and Winter 2024
DOI: 10.71586/jal.2024.11181190893

of relying on generalized and empirically unsupported descriptors, it is more effective to develop task-specific, empirically derived, and diagnostically oriented items (e.g., Khamboonruang, 2020; Kim, 2019; Lukácsi, 2021; Safari & Ahmadi, 2023; Zou et al., 2024). Such assessment tools should be informed by research and tailored to specific language use contexts, providing a more precise assessment of key components of language competence.

These challenges have led researchers to adopt a variety of methodological approaches aimed at enhancing the reliability, validity, and diagnostic precision of writing assessment scales. During the development phase, think-aloud protocols and expert feedback are frequently used to refine scale descriptors, ensuring they accurately reflect real-world assessment behaviors and provide meaningful diagnostic insights (e.g., Kim, 2019; Lukácsi, 2021; Ma et al., 2022; Wagner, 2015). Once developed, these scales undergo rigorous statistical testing (e.g., Cohen's Kappa, Pearson correlation, intra-rater reliability coefficients) to confirm their reliability and validity across different contexts and rater groups, establishing a consistent framework for scoring.

For instance, Wang and Xie (2022) employed inter-rater and intra-rater reliability coefficients to verify the marking consistency of the course assessment rubric, applying methods outlined by Cohen (2017, as cited in Wang & Xie, 2022). After confirming sufficient reliability, the researchers used the newly developed diagnostic rubric to assess student responses and analyze writing performance. Similarly, Safari and Ahmadi (2023) used correlation analyses to examine the relationship between diagnostic checklist scores and those awarded by TOEFL iBT's holistic rubric, alongside measuring both consistency and consensus estimates to assess inter-rater agreement across descriptors.

Despite these methodological efforts, challenges remain in ensuring that rating scales align with how raters naturally evaluate writing performance. Knoch, Deygers, and Khamboonruang (2021) underscore the centrality of rating processes in rater-mediated performance assessment, emphasizing that "rating scales form a key part of that link between a performance and a claim (or score)" (p. 2). The authors go on to argue that while "The Standards for Educational and Psychological Testing" (AERA et al., 2014, as cited in Knoch et al., 2021) and other publications stress the importance of well-defined scoring criteria, "little information [is] available on how rating scales (also referred to as scoring criteria, or scoring rubrics) are developed [and] what sources may influence the way the rating scale reflects the wider test construct" (p. 3). This lack of clarity can lead

**The Journal of English Language Pedagogy and Practice**
Vol. 17, No.35, Autumn and Winter 2024
DOI: 10.71586/jal.2024.11181190893

to discrepancies between descriptors and rater interpretations, hindering reliability in assessment outcomes.

In response to these challenges, the present study sought to develop a diagnostic rating scale specifically tailored for the EFL academic context, emphasizing the need for empirically derived and diagnostically oriented descriptors. To address the potential misalignment between scale criteria and raters' evaluation processes, the study employed a combination of think-aloud protocols, expert feedback, and cross-referencing with established rating scales and theories of writing during the development phase, ensuring that the descriptors were grounded in real-world assessment behaviors. Additionally, carefully designed training sessions and clear, comprehensive guidelines were provided to raters to promote consistency and reliability in scoring. This multi-step approach aimed to create a robust, practical tool capable of delivering precise, fine-grained feedback, thereby overcoming the limitations of traditional assessment scales in diagnosing writing performance. Guided by four specific research questions, the study highlights the importance of contextually relevant and empirically grounded assessment tools in fostering effective writing instruction for EFL settings.

1. What are the key descriptors of the diagnostic scale tailored to EFL learners?
2. To what extent do the results from the Content Validity Index (CVI) analysis support the content validity of the scale in assessing EFL learners' writing proficiency?
3. How consistent are the results of the inter-rater reliability analysis when the scale is applied to different writing samples?
4. To what extent do the scores awarded by the raters using the diagnostic scale correlate with the original scores provided by the instructor?

## Method

### Participants

The selection of the participants in the study followed qualitative research techniques which "is purposeful (or intentional) sampling through the selection of a sample of participants who can best help the researcher understand the central phenomenon being explored" (Creswell, 2022, p. 88). Therefore, the participants were selected based on careful consideration of several criteria, given the qualitative nature of the initial stage of the scale development.

First, their expertise, background knowledge, and familiarity with academic writing conventions were critical for reliable and meaningful data collection. Second, their experience teaching EFL writing and high-stakes courses (e.g.,

**The Journal of English Language Pedagogy and Practice**
Vol. 17, No.35, Autumn and Winter 2024
DOI: 10.71586/jal.2024.11181190893

IELTS or TOEFL) were essential for ensuring consistency in essay rating (i.e., by raters) and writing performance (i.e., by writers). Third, for the think-aloud protocols, it was necessary to involve raters and writers capable of reflecting on their respective processes—assessment for raters and composition for writers—with a deep understanding of academic writing criteria. Lastly, participants with advanced qualifications and practical experience were necessary to provide valid, contextually relevant feedback during the review and evaluation of the scale.

Four groups of participants were involved in different stages of this study. The first two groups participated in the think-aloud protocols while the latter two evaluated the developed diagnostic writing scale. All participants were affiliated with accredited universities in Tehran where they held roles related to EFL instruction and research. They all had at least ten years of experience teaching EFL writing and international language proficiency exams (e.g., IELTS, TOEFL, or similar). Further details about their tasks and involvement are explained in the Procedure section. The participants were 23 teachers including:

1. five raters (four PhD graduates and one PhD candidate in TEFL) who participated in the think-aloud protocols;
2. five writers (three PhD candidates and two MA graduates in TEFL) who participated in the think-aloud protocols;
3. four writing experts (one PhD graduate and three PhD candidates in TEFL) who reviewed and provided feedback on the scale; and
4. nine raters (two PhD graduates, five PhD candidates, and two MA graduates in TEFL) who evaluated the diagnostic scale.

## Materials and Instruments

The following materials and instruments were used during the initial stage of the study to capture a comprehensive understanding of the rating and writing processes among EFL learners and educators:

1. IELTS Task 2 Writing Samples: 20 essays (5 essays x 4 prompts) representing different proficiency levels were obtained from the students preparing for the IELTS Task 2 writing at a language institute in Tehran and used for the five raters' think-aloud sessions. IELTS Task 2 requires test-takers to write a 250-word essay on academic topics, such as education, social issues, and technology, within a 40-minute time frame. The task assesses skills relevant to real-world academic contexts, including discussion and opinion essays. The four prompts were selected to reflect these common themes and examined to ensure comparable difficulty levels, with scores ranging from 3.5 to 8 on the IELTS 9-band

**The Journal of English Language Pedagogy and Practice**
Vol. 17, No.35, Autumn and Winter 2024
DOI: 10.71586/jal.2024.11181190893

scale (1 = non-user, 9 = expert user), providing a balanced representation of diverse writing abilities.

2. IELTS Task 2 Writing Prompt: An IELTS task 2 prompt was presented to five participants to verbalize their thoughts while answering the question in 45 minutes.
3. Cambridge Online English Placement Test (CEPT): Designed to assess proficiency across the Common European Framework of Reference for Languages (CEFR) levels, the CEPT offers a quick and objective method to gauge language proficiency. Administered as a pretest, this 15-minute online assessment verified that all writer participants met the study's advanced proficiency requirement.
4. Essay Writing Task: To further confirm the participants' academic writing skills, the writers completed a 40-minute essay on a specified prompt. A minimum equivalent IELTS writing score of 8 was required, ensuring that the participants demonstrated high-level writing ability.
5. Think-Aloud Verbal Protocol and Guidelines: The writers and raters expressed their thoughts aloud during their respective sessions, directed by standardized guidelines to ensure consistency in verbalization and feedback. These guidelines were meticulously developed and strictly followed, crucial for minimizing variability in data collection and enhancing the reliability of the insights gathered. The instructions clearly outlined how participants should articulate their thoughts and reasoning comprehensively, thereby ensuring accurate capture of their cognitive processes.
6. Background Questionnaire: Comprised of demographic information, assessment practices, and an evaluation of EFL writing skills, the participants rated the importance of the criteria derived from widely-used writing rubrics.
7. Open-ended Interviews: Post-think-aloud interviews gathered the participants' insights on the think-aloud procedure and academic writing.

The following materials and instruments were utilized during the pilot testing phase to validate the effectiveness and reliability of the newly developed diagnostic scale:

8. Content Validity Index (CVI) Questionnaire assessed the content validity of the 21 descriptors through expert feedback. The questionnaire consisted of three sections: (a) demographic information, including teaching and assessment experience; (b) a 4-point Likert scale, where raters evaluated

**The Journal of English Language Pedagogy and Practice**
Vol. 17, No.35, Autumn and Winter 2024
DOI: 10.71586/jal.2024.11181190893

each descriptor from lowest (e.g., not clear) to highest (e.g., very clear); and (c) two open-ended questions to gather suggestions or report any issues.

9. Pilot Phase Essays: Five additional IELTS Task 2 essays were used in the pilot phase to validate the diagnostic scale. These essays were selected using the same protocols outlined earlier for the IELTS Task 2 writing samples, ensuring consistency across both stages.

## Procedure

The primary aim of this study was to identify and define key descriptors for a diagnostic academic writing scale tailored to EFL learners. Designed to provide targeted, diagnostic feedback, the scale seeks to support both instructional practices and learner progress. To achieve this, a mixed-methods approach was employed, incorporating think-aloud verbal reports, interviews, and questionnaires to capture the cognitive processes and perspectives of five experienced raters as they evaluated 10 IELTS Task 2 essays and five EFL teachers as they completed an IELTS Task 2 essay.

The essays for the think-aloud sessions were carefully selected to cover a wide range of writing behaviors and proficiency levels. This ensured a comprehensive analysis of the criteria that raters use in assessing writing. The selection aimed to capture varied feedback and rating practices across different types of writing prompts, which are essential for developing a scale that is responsive to diverse academic writing needs. This preparatory stage, focused on understanding the raters' and writers' interactions with various essay topics, was critical for refining the descriptors before testing their reliability and validity. This methodology aligns with recommendations by Knoch (2009) and Weigle (2002) to incorporate diverse topics and proficiency levels early in the development process to enhance the scale's generalizability.

Before the actual ratings began, the researcher conducted one-on-one training sessions with each participant, providing clear guidelines on the think-aloud procedure. The raters were instructed to verbalize their thoughts while evaluating each essay and to assign a final score based on their expertise in EFL academic essay writing, with scores ultimately aligned to IELTS Task 2 band descriptors (i.e., ranging from 1 to 9) to facilitate correlation analysis with the instructor's original scores. Similarly, writers were instructed to verbalize their thoughts while responding to an IELTS Task 2 prompt.

The think-aloud protocols for the raters followed a concurrent verbalization approach, with the participants verbalizing their thoughts in real-time while rating

**The Journal of English Language Pedagogy and Practice**
Vol. 17, No.35, Autumn and Winter 2024
DOI: 10.71586/jal.2024.11181190893

the essays. In contrast, three writers opted for retrospective protocols due to the cognitive load of writing and speaking simultaneously, while two writers comfortably followed the concurrent approach, providing comments during the composition process.

After completing the task, the participants took part in a retrospective interview session to reflect on the think-aloud process, difficulties encountered, criteria used for assessment, and their perspectives on the EFL academic writing features (e.g., "What do you usually focus on when providing feedback on EFL academic essays?" and "What do you usually focus on when composing an academic essay?"). These interviews offered deeper insights into the participants' cognitive processes and their perceptions of the writing criteria. Each think-aloud session lasted approximately 3 to 4 hours, followed by an additional 2-hour interview and questionnaire session. The entire data collection process was conducted online.

The think-aloud sessions and interviews were audio-recorded, then transcribed verbatim to capture both verbal and non-verbal cues, offering insights into the participants' assessment or composition processes. The transcripts were formatted for coding, with clear timestamps and speaker identification, then systematically organized for analysis. To ensure the data reliability and consistency, correlation analysis was conducted for the five raters' scores. These evaluations confirmed that the scoring patterns were consistent before thematic analysis began. The writers' proficiency was verified prior to the think-aloud sessions through the perfect scores on the Cambridge Online English Placement Test (Cambridge Assessment English, n.d.) and a minimum score of 8 on an IELTS Task 2 essay prompt.

Thematic analysis, "a method for identifying themes in qualitative data" (Terry, Hayfield, Clarke, & Braun, 2017, p. 17), was the primary methodology used to extract the key descriptors. Following the transcription of think-aloud sessions and interviews, as well as the collection of questionnaire data, the analysis proceeded through six structured steps as described by Terry et al. (2017): familiarizing with the data, generating initial codes, constructing themes, reviewing themes, defining and naming themes, and producing the report.

The familiarization phase involved thoroughly reading and re-reading the transcripts and questionnaire responses to gain an in-depth understanding of the participants' perspectives. During this stage, notes were taken to identify the preliminary patterns related to the writing features such as organization, coherence, and grammar. Next, in the initial coding phase, an inductive, open

**The Journal of English Language Pedagogy and Practice**
Vol. 17, No.35, Autumn and Winter 2024
DOI: 10.71586/jal.2024.11181190893

coding approach was employed to capture both explicit and underlying aspects of writing performance described in verbal and written feedback. These codes represented distinct writing features mentioned by the participants, such as "topic sentences," "supporting evidence," and "sentence variety."

During the theme construction stage, similar or related codes were grouped to form broader categories, representing potential themes. In total, 75 distinct codes were identified, reflecting various aspects of EFL academic writing. These codes were systematically reviewed and refined in iterative cycles to ensure conceptual clarity and alignment with established writing theories (e.g., Cumming, 2001; Grabe & Kaplan, 1996; Hayes, 1996; Sasaki & Hirose, 1996) and diagnostic frameworks (e.g., He et al., 2021; Khamboonruang, 2020; Kim, 2019; Safari & Ahmadi, 2023; Shi, Ma, Du, & Gao, 2024).

The codes with overlapping meanings were merged while those lacking sufficient evidence were discarded or revised, condensing them into 44 themes. For example, the codes related to logical structure, transitions, and paragraph development were grouped under the theme of 'organization,' capturing how ideas were structured and presented within essays. In contrast, a code related to isolated errors in prepositions, which appeared infrequently and lacked broader relevance, was discarded to streamline the theme of 'grammatical range and accuracy.'

In the defining and naming phase, the 44 themes were systematically refined through a combination of methods: calculating the percentage of agreement among the participants, collecting feedback from four writing experts, and conducting additional thematic reviews. This iterative process led to a final set of 21 themes, representing critical components of EFL academic writing assessment, such as coherence, lexical variety, grammatical range and accuracy, and argument development. Finally, in the production step, the themes were transformed into short, simple, and unambiguous sentences, forming the basis of the diagnostic writing scale descriptors, following a process similar to that employed by Kim (2019) and Safari & Ahmadi (2023).

Following the initial scale development, the descriptors underwent content validation to ensure they captured essential elements of writing assessment. Romero Jeldres, Díaz Costa, and Faouzi (2023) emphasize the importance of relevance and representativeness for measurement instruments, stating, "measuring the correspondence between the content of the items and the evaluated content is very relevant for evaluating content validity" (Romero Jeldres et al., 2023, p. 1). In line with this perspective, a Content Validity Index (CVI, Lawshe,

**The Journal of English Language Pedagogy and Practice**
Vol. 17, No.35, Autumn and Winter 2024
DOI: 10.71586/jal.2024.11181190893

1975) questionnaire, consisting of six dimensions—clarity, comprehensiveness, importance, relevance, redundancy, and simplicity—was developed to assess the diagnostic descriptors. These dimensions were selected to ensure both theoretical soundness and practical applicability.

The development of this questionnaire was guided by three key considerations: (a) a review of existing literature on content validity (e.g., Romero Jeldres et al., 2023; Masuwai, Zulkifli, & Hamzah, 2024; Mukminin, Habibi, Muhaimin, & Hidayat, 2023) and diagnostic writing frameworks (e.g., Kim, 2019; Shi et al., 2024), (b) feedback from a panel of four writing experts who were also actively involved in refining the diagnostic scale, and (c) contextual adjustments to reflect the diagnostic needs of the study. Additionally, nine raters participated in the evaluation process, consistent with recommendations that panels of five to ten experts optimize both reliability and feasibility (Baghestani, Ahmadi, Tanha, & Meshkat, 2019; Romero Jeldres et al., 2023; Lawshe, 1975; Lynn, 1986).

The feedback from the CVI assessment led to minor adjustments in the wording of the descriptors to improve clarity. A pilot test was then conducted with eight of the nine panel experts (one expert was unable to participate) to assess the reliability and consistency of the raters when applying the refined descriptors. The pilot aimed to ensure the scale's functionality in practice, particularly in identifying and addressing any ambiguities in the descriptors. For this phase, a set of five essays—previously rated by the course instructor using IELTS Task 2 band descriptors—was used. The raters applied the diagnostic scale to these essays, enabling an evaluation of both inter-rater reliability and concurrent criterion-related validity.

In this study, intraclass correlation coefficient (ICC) served to quantify the inter-rater reliability by evaluating how consistently the raters applied the diagnostic scale across different essays and scoring criteria. The ICC is a reliability index that reflects both the degree of correlation and agreement between measurements (Koo & Li, 2016). It measures the proportion of total variance attributable to differences in the target trait—such as writing ability—versus unwanted variance from inconsistencies between raters (Liljequist, Elfving, & Skavberg Roaldsen, 2019). Criterion-related validity was assessed through the correlation between the raters' scores and the instructor's original scores, reflecting its definition as the relationship between test performance and an external benchmark. Concurrent validity, in particular, involves comparing scores from a test and another measure of the same ability, collected simultaneously (Weir, Chan, & Nakatsuhara, 2013).

**The Journal of English Language Pedagogy and Practice**
Vol. 17, No.35, Autumn and Winter 2024
DOI: 10.71586/jal.2024.11181190893

## Research Design

The present study aims to answer four research questions examining the diagnostic assessment of Iranian higher education students' academic writing proficiency. The overall research design of the study is the mixed-methods approach, incorporating the techniques of both qualitative and quantitative methods of inquiry within a single study (Creswell, 2022). Mixed methods design strives to integrate multiple methods with multiple perspectives for gathering evidence in order to reduce inherent biases and limitations of a single method (Greene, Caracelli, & Graham, 1989). The gathered evidence from various data sources including qualitative and quantitative methods were used to confirm the findings and provide a logical response to the research questions. The above sections have summarized the participants, materials and instruments, and data collection and analysis procedures associated with each stage of the study.

## Data Analysis

For the first research question, the reliability and consistency of the data collected from the raters were evaluated using the Intraclass Correlation Coefficient (ICC) and Pearson correlation coefficient, analyzed through SPSS version 20. After confirming the data reliability, a thematic analysis was conducted, supported by frequency counts to quantify agreement and extract key themes. The second research question was addressed using a Content Validity Index (CVI) to assess various dimensions of the descriptors based on the expert feedback. Following the construction and content validation of the scale, a pilot test was conducted to evaluate its functionality. The third research question focused on measuring inter-rater reliability through ICC, ensuring that the scale produced consistent scores. Lastly, Pearson correlation was used to address the fourth question by examining concurrent criterion-related validity, comparing the diagnostic scale's scores with the original proficiency scores.

# Results

## Addressing Research Question 1

To ensure the reliability and validity of the verbal data collected from the raters and writers, several preliminary assessments were conducted. For the writers, the Cambridge Online English Placement Test was administered as a pretest to confirm their advanced proficiency in English, with each participant scoring 100. The five writers also completed a 40-minute essay on a specified prompt,

**The Journal of English Language Pedagogy and Practice**
Vol. 17, No.35, Autumn and Winter 2024
DOI: 10.71586/jal.2024.11181190893

achieving an IELTS-equivalent score of 8 or above, thus, confirming their eligibility for the study.

For the raters, the consistency of scoring patterns was measured using ICC (Fraga-Viñas, 2022; Liljequist et al., 2019) and Pearson *r* (Jin, 2023) to ensure reliable data collection before proceeding with thematic analysis (Terry et al., 2017). The inter-rater reliability was assessed by having the raters evaluate two separate batches of the essays written on four different prompts, providing diversity in proficiency levels and topic coverage.

Upon calculating the inter-rater reliability for two raters who assessed the first batch of 10 essays, a high level of agreement was observed. The Single Measures ICC was 0.824, indicating strong consistency between the raters' scores, while the Average Measures ICC was 0.904, demonstrating excellent reliability. An F-Test statistic ($F = 10.378$, $p = 0.001$) confirmed that this consistency was statistically significant. For the second batch of the essays assessed by three raters, similarly high inter-rater reliability was identified. The Single Measures ICC reached 0.933 (95% CI: 0.820 to 0.981), while the Average Measures ICC rose to 0.977 (95% CI: 0.932 to 0.994), confirming excellent reliability. This was further supported by the F-Test ($F = 42.841$, $p < 0.001$).

A Pearson correlation analysis was conducted to examine the consistency between the scores from the five raters and the course instructor's original scores, with correlation values ranging from 0.818 to 1.000 (Table 1). These results indicate strong to perfect alignment with the established benchmark scores, providing evidence of scoring consistency during the assessment process. The perfect and near-perfect correlations observed for Rater 4 and Rater 5 can be explained by the fact that both raters were professional IELTS instructors and examiners while the essays evaluated were originally written by the students preparing for the IELTS exam. At this stage of the study, the raters applied their own established criteria for writing assessment, which closely aligned with the IELTS rubric in the case of these two raters.

Table 1
*Correlation between Raters' Scores and Instructor's Original Scores*

| Rater | Pearson Correlation | p-value |
|---|---|---|
| Rater 1 | 0.892 | 0.001 |
| Rater 2 | 0.818 | 0.004 |
| Rater 3 | 0.978 | <0.001 |
| Rater 4 | 1.000 | <0.001 |
| Rater 5 | 0.930 | <0.001 |

**The Journal of English Language Pedagogy and Practice**
Vol. 17, No.35, Autumn and Winter 2024
DOI: 10.71586/jal.2024.11181190893

*Note: To maintain confidentiality, the raters were assigned numerical labels rather than using their names.*

After confirming the reliability of the participants' responses, thematic analysis was conducted to extract the key descriptors essential for assessing EFL writing proficiency. The analysis began with 75 initial codes, which were systematically refined through the iterative cycles of review, expert consultation, and alignment with established writing theories and diagnostic frameworks. These codes were grouped into 44 preliminary themes, with overlapping themes consolidated and less relevant themes rephrased or removed. For example, themes like "errors in the use of prepositions" and "errors in the use of articles" were merged into "correct use of function words (e.g., articles, modals, prepositions, pronouns)," while less contributory themes, such as "sufficient boosting and hedging," were excluded.

This iterative refinement, supported by frequency counts and expert feedback, resulted in the creation of 21 final descriptors derived from the initial themes. These descriptors represent the essential components of EFL academic writing, with participant agreement reflected through frequency and percentage data. Table 2 highlights key descriptors, such as "the essay has a clear thesis statement" and "the essay presents well-developed ideas," which received unanimous agreement (100%) as critical to effective academic writing. The highest priority was assigned to descriptors related to argument development and organization, while aspects of sentence structure, such as avoiding fragmented and run-on sentences, received comparatively lower emphasis (65.21%).

Table 2
*Frequency and Percentage of Participant Agreement on 21 Descriptors of the Diagnostic Scale*

| Descriptor | Participants (10) | Experts (4) | Raters (9) | Total Frequency (*f*) | Percentage (%) |
|---|---|---|---|---|---|
| D1 | 5 | 4 | 9 | 18 | 78.26% |
| D2 | 10 | 4 | 9 | 23 | 100% |
| D3 | 10 | 4 | 9 | 23 | 100% |
| D4 | 9 | 4 | 9 | 22 | 95.65% |
| D5 | 9 | 4 | 9 | 22 | 95.65% |
| D6 | 8 | 3 | 9 | 20 | 86.95% |
| D7 | 9 | 4 | 9 | 22 | 95.65% |
| D8 | 10 | 4 | 9 | 23 | 100% |
| D9 | 9 | 4 | 9 | 22 | 95.65% |
| D10 | 9 | 4 | 7 | 20 | 86.95% |

**The Journal of English Language Pedagogy and Practice**
Vol. 17, No.35, Autumn and Winter 2024
DOI: 10.71586/jal.2024.11181190893

| | | | | | |
|---|---|---|---|---|---|
| D11 | 7 | 3 | 7 | 17 | 73.91% |
| D12 | 7 | 3 | 9 | 18 | 78.26% |
| D13 | 9 | 4 | 9 | 22 | 95.65% |
| D14 | 9 | 4 | 9 | 22 | 95.65% |
| D15 | 7 | 3 | 9 | 19 | 82.60% |
| D16 | 7 | 4 | 9 | 20 | 86.95% |
| D17 | 8 | 4 | 9 | 21 | 91.30% |
| D18 | 5 | 4 | 9 | 18 | 78.26% |
| D19 | 8 | 4 | 9 | 21 | 91.30% |
| D20 | 8 | 4 | 9 | 21 | 91.30% |
| D21 | 8 | 4 | 9 | 21 | 91.30% |

*Note: Participants include five raters and five writers from the initial data collection. Experts refer to the four writing specialists consulted during scale development. Raters indicate the nine experts involved in the pilot testing phase.*

## Addressing Research Question 2

In this study, the content validity was assessed to ensure that the diagnostic writing scale comprehensively represents the essential components of the academic writing proficiency for EFL learners. The Content Validity Ratio (CVR), which quantifies the agreement among experts on an item for accurately representing a content domain (Masuwai et al., 2024), was computed for each of the 21 descriptors across six dimensions: clarity, comprehensiveness, importance, relevance, redundancy, and simplicity. This calculation followed Lawshe's (1975) method, identifying the proportion of experts who rated each descriptor as valid (i.e., assigning a score of "3" or "4" on a 4-point Likert scale). The scale items were evaluated using a questionnaire designed with definitions for each dimension. For example, clarity referred to whether the descriptor was understandable and unambiguous. The experts rated each item on levels ranging from "not clear" to "very clear", with similar rating options applied to other dimensions. The formula used for CVR calculations accounted for the number of experts rating the item as valid ($n_{valid}$) and the total number of experts ($N$ = 9), with a critical CVR value of 0.653 at the 0.05 significance level (Wilson, Pan, & Schumsky, 2012). All descriptors exceeding this threshold were considered valid for their respective dimensions, consistent with the findings from other studies (e.g., Baghestani et al., 2019; Romero Jeldres et al., 2023; Masuwai et al., 2024).

$$\text{CVR} = \frac{n_{valid} - (N/2)}{N/2}$$

**The Journal of English Language Pedagogy and Practice**
Vol. 17, No.35, Autumn and Winter 2024
DOI: 10.71586/jal.2024.11181190893

Furthermore, the Content Validity Index (CVI), which evaluates how well an instrument includes appropriate items to represent the intended construct, was calculated (Mukminin, 2023; Lynn, 1986). Two types of CVI were used: the Item-level Content Validity Index (I-CVI), which measured the proportion of the experts rating each item as valid, and the Scale-level Content Validity Index (S-CVI), which was derived through two methods: the average I-CVI across all items (S-CVI/Ave) and the proportion of items achieving perfect agreement among raters (S-CVI/UA) (Shi, Mo, & Sun, 2012).

For clarity, most descriptors achieved a CVR of 1.00, resulting in an S-CVI/Ave of 0.99, indicating that the descriptors were well-understood and unambiguous to the experts. In terms of comprehensiveness, descriptors scored similarly high, with only item 12 receiving a slightly lower CVR of 0.889. This still exceeded the threshold, confirming that all descriptors covered the necessary aspects of academic writing. For importance, all items received high ratings, with a CVR and S-CVI/UA of 1.00, reflecting agreement on the critical role of each descriptor in assessing EFL writing. The relevance dimension, evaluating the descriptors' applicability to real-world academic writing, showed an overall CVR of 0.98, with item 12 again scoring slightly lower at 0.889 but still surpassing the threshold. Redundancy checks confirmed that descriptors were distinct, with a CVR of 1.00 and an S-CVI/UA of 0.97, affirming no overlap between items. Simplicity, which assessed the ease of application and interpretation, also received excellent ratings, with a CVR and S-CVI/UA of 1.00 across all items.

Overall, the scale's S-CVI/Ave was 0.994, reflecting a high level of expert agreement on the descriptors' content representativeness. The S-CVI/UA was 0.97, indicating that the majority of the descriptors achieved perfect agreement. These high ratings across all dimensions confirmed the robustness of the descriptors and validated their relevance and representativeness for academic writing assessment. As noted by Romero Jeldres et al. (2023), high CVI ratings "allow the maintenance of a vision of the relevance and representativeness of the items" (p. 4). Additionally, no items required removal or major refinement, given consistently high CVR values (Baghestani et al., 2019). However, item 12, though passing the threshold (Romero Jeldres et al., 2023; Wilson et al., 2012), underwent slight rewording in consultation with experts to enhance its clarity and relevance.

## Addressing Research Question 3

Following the content validity assessment, a reliability analysis was conducted to determine the consistency and agreement among the raters using the diagnostic

**The Journal of English Language Pedagogy and Practice**
Vol. 17, No.35, Autumn and Winter 2024
DOI: 10.71586/jal.2024.11181190893

academic writing scale. Given the multiple raters and item-based scoring structure, the Intraclass Correlation Coefficient (ICC) was selected as the most appropriate measure of inter-rater reliability (Liljequist et al., 2019). A two-way random effects model (ICC(2,1)) was used to assess reliability, treating both raters and essays as random samples from larger populations (Koo & Li, 2016; Liljequist et al., 2019; Shrout & Fleiss, 1979). This model evaluated consistency across the raters while assuming random variability for both sources. The ICC results, presented in Table 3, indicate both Single Measures and Average Measures ICC values.

Table 3
*Intraclass Correlation Coefficient (ICC) Analysis for Inter-Rater Reliability*

| Statistic | Value | 95% Confidence Interval | F-Test (df1, df2) | *p*-value |
|---|---|---|---|---|
| Single Measures ICC | 0.577 | 0.258 to 0.925 | F = 13.271 (4, 32) | < 0.001 |
| Average Measures ICC | 0.925 | 0.757 to 0.991 | F = 13.271 (4, 32) | < 0.001 |

The Single Measures ICC of 0.577, in Table 3, suggests moderate reliability when considering individual raters' scores, with a 95% confidence interval ranging from 0.258 to 0.925. These values fall within the thresholds provided by previous studies and indicate acceptable but moderate reliability (e.g., Fraga-Viñas, 2022; Koo & Li, 2016; Liljequist et al., 2019). The Average Measures ICC of 0.925, with a 95% confidence interval of 0.757 to 0.991, reflects excellent reliability when averaging scores across all raters. An F-test value of 13.271 ($p < 0.001$) supports these results, indicating that the observed variability in scores was significantly greater than what would be expected by chance.

## Addressing Research Question 4

To address Research Question 4, a Pearson correlation analysis was conducted to examine the relationship between the scores awarded by the raters using the diagnostic writing scale and the original scores provided by the instructor based on IELTS Task 2 band descriptors. This analysis aimed to assess the concurrent validity of the scale by evaluating the alignment between the diagnostic scale ratings and the instructor's expert evaluations. The concurrent validity, in this context, refers to how well the test scores correlate with an established external measure when both are assessed simultaneously (Weir et al., 2013). The Pearson correlation coefficients between the original instructor scores and the scores given by each of the eight diagnostic scale raters are presented in Table 4 below.

**The Journal of English Language Pedagogy and Practice**
Vol. 17, No.35, Autumn and Winter 2024
DOI: 10.71586/jal.2024.11181190893

Table 4
*Correlation Coefficients between Instructor Ratings and Diagnostic Scale Ratings*

| Rater | Pearson Correlation (r) | *p*-value | Interpretation |
|---|---|---|---|
| Rater1 | 0.916* | 0.029 | Strong Positive Correlation (Significant) |
| Rater2 | 0.899* | 0.038 | Strong Positive Correlation (Significant) |
| Rater3 | 0.974** | 0.005 | Very Strong Positive Correlation (Highly Significant) |
| Rater4 | 0.977** | 0.004 | Very Strong Positive Correlation (Highly Significant) |
| Rater5 | 0.946* | 0.015 | Strong Positive Correlation (Significant) |
| Rater6 | 1.000** | <0.001 | Perfect Correlation (Highly Significant) |
| Rater7 | 0.900* | 0.037 | Strong Positive Correlation (Significant) |
| Rater8 | 0.900* | 0.037 | Strong Positive Correlation (Significant) |

*Note: Correlations marked with * are significant at the 0.05 level (2-tailed), and those marked with ** are significant at the 0.01 level (2-tailed).*

The Pearson correlation coefficients, ranging from 0.899 to 1.000, demonstrate strong to very strong positive correlations. These results indicate that the scores assigned by the raters using the diagnostic scale closely correspond to the instructor's original scores. This correspondence affirms the scale's ability to reliably measure the key aspects of the academic writing proficiency. Similar findings have been reported in other validation studies (e.g., Lin, 2024; Safari & Ahmadi, 2023). All correlations were statistically significant, with several achieving significance at the 0.01 level. Raters 3, 4, and 6 displayed very strong positive correlations ($r$ = 0.974, 0.977, and 1.000, respectively) with highly significant results ($p \leq 0.005$). Other raters, including Rater1, Rater2, Rater5, and Rater7, also showed strong positive correlations ($p < 0.05$), further supporting the scale's consistency and diagnostic effectiveness.

## Discussion

To address the first research question, the study sought to identify the key descriptors for a diagnostic-oriented rating scale of academic writing, specifically tailored for EFL learners. The aim was to develop a scale that goes beyond traditional, generalized rubrics by offering more detailed, diagnostic feedback on learners' writing skills. This focus is consistent with the existing research which highlights the importance of creating empirically-based diagnostic tools that can precisely capture and assess multiple facets of writing (He et al., 2021; Kim, 2019; Safari & Ahmadi, 2023; Shi et al., 2024; Wagner, 2015).

The development of the diagnostic-oriented rating scale employed a combination of qualitative methodologies, including think-aloud protocols and

**The Journal of English Language Pedagogy and Practice**
Vol. 17, No.35, Autumn and Winter 2024
DOI: 10.71586/jal.2024.11181190893

expert feedback sessions. Experienced EFL teachers participated in the verbalization process, serving both as raters and writers, by articulating the key criteria they considered essential for evaluating academic writing. This dual perspective enabled the identification of critical elements reflecting real-world assessment behaviors from both the teaching and learning standpoints. Insights were gained into the criteria raters prioritize when evaluating various aspects of writing and how writers recognize and define the essential components of effective academic writing. This comprehensive approach facilitated a deeper understanding of essential skills in EFL academic writing, ensuring that the descriptors were aligned with the practical realities of both assessment and instruction.

Beyond empirical data collection, the descriptors were cross-referenced with established theories of writing (e.g., Cumming, 2001; Grabe & Kaplan, 1996; Hayes, 1996; Sasaki & Hirose, 1996) and existing diagnostic frameworks (e.g., He et al., 2021; Khamboonruang, 2020; Kim, 2019; Lu et al., 2021; Safari & Ahmadi, 2023; Shi et al., 2024), ensuring their theoretical soundness and practical relevance. This process aimed to generate descriptors that capture essential components of EFL academic writing proficiency, grounded in both theoretical constructs and real-world assessment needs.

The final set of descriptors was refined to address three core areas of academic writing proficiency: content fulfillment, organizational knowledge, and language usage. Unlike previous scales that might rely on general, less specific evaluation criteria, the diagnostic scale developed in this study provides detailed descriptors for each writing skill. For example, descriptors under content fulfillment focus on thesis clarity and the adequate development of ideas, emphasizing the importance of content in writing assessment (Kim, 2019; Shi et al., 2024; Wagner, 2015). Similarly, descriptors under organizational knowledge assess the logical arrangement of ideas, effective use of cohesive devices, and clear paragraph structure (Kim, 2019; Shi et al., 2024), while those under language usage emphasize precision in vocabulary, grammatical range, and appropriate punctuation (He et al., 2021).

Overall, the development of this scale directly responded to the need for a diagnostic tool that could provide comprehensive feedback across multiple dimensions of writing proficiency. Through a robust empirical process, the study successfully identified a set of 21 descriptors, offering nuanced, diagnostic feedback to guide instructional strategies and support learner development in EFL academic writing.

**The Journal of English Language Pedagogy and Practice**
Vol. 17, No.35, Autumn and Winter 2024
DOI: 10.71586/jal.2024.11181190893

For the second research question, the study examined the content validity of the diagnostic scale, to ensure that it accurately and comprehensively represents the construct it intends to assess, making it both theoretically grounded and applicable in practice (Masuwai et al., 2024; Polit & Beck, 2006; Rubio, Berg-Weger, Tebb, Lee, & Rauch, 2003). By establishing content validity early in development, researchers can create more reliable tools that minimize the need for later revisions, streamlining further validation phases such as construct validity and reliability testing (Rubio et al., 2003). As the authors aptly assert, "measures that have established content validity would need fewer revisions in the evaluation phase" (p. 95).

A content validity index (CVI) analysis was conducted with a panel of nine academic writing specialists, evaluating the 21 descriptors across six key dimensions: clarity, comprehensiveness, importance, relevance, redundancy, and simplicity. This emphasis on content validation is consistent with previous studies (Lynn, 1986; Polit & Beck, 2006; Romero Jeldres et al., 2023), which emphasize rigorous expert review and the CVI process to ensure that a scale's items represent the target construct. The results confirmed that the descriptors adequately represent the domain of content, as reflected in high CVI and CVR scores. For example, the descriptors were found to be clear and unambiguous, comprehensive in covering critical writing aspects, and highly relevant to EFL academic writing contexts. This robust evaluation led to only minor adjustments in wording, reinforcing the strength of the scale development process.

Ultimately, the high ratings across all dimensions and a final S-CVI/Ave of 0.994 confirmed the scale's representativeness and relevance as a diagnostic tool for EFL academic writing assessment. However, it should be noted that, 'experts' feedback is subjective; thus, the study is subjected to bias that may exist among the experts. In addition, this type of study does not eliminate the need for additional psychometric testing, which is critical for the development of a measure" (Rubio et al., 2003, pp. 95-96).

To mitigate potential bias, initiatives were undertaken to ensure fair evaluation, including the selection of experts with diverse backgrounds and experiences in writing and EFL assessment, the use of detailed criteria to guide evaluations, and independent assessments conducted without influence from others. Nevertheless, the current study proceeded with a pilot test to further examine the scale's other psychometric properties.

In evaluating the third research question, the study analyzed the consistency in diagnostic scale ratings through the Intraclass Correlation Coefficient (ICC)

**The Journal of English Language Pedagogy and Practice**
Vol. 17, No.35, Autumn and Winter 2024
DOI: 10.71586/jal.2024.11181190893

analysis, which served as a robust measure of inter-rater reliability across different writing samples. The two-way random effects model (ICC2,1) was used, emphasizing the importance of consistency over absolute agreement to focus on the relative ranking of essays by different raters.

The Single Measures ICC of 0.577 suggested moderate reliability at the level of individual raters. This moderate reliability, while reflective of the challenges inherent in subjective evaluations, underscores the importance of measures that can support rater consistency. Variability in individual ratings is common in subjective assessments, especially in writing evaluation, where individual judgment can introduce slight differences. Kim (2019) contends, "it is extremely difficult for raters to achieve substantial agreement on writing assessments, possibly because of the inherently subjective nature of the task" (p. 916).

Despite employing well-trained professional raters, both Knoch (2009) and Kim (2019) observed only moderate agreement levels. Knoch's study demonstrated moderate agreement, with alignment ranging from 36.1% to 61.9% on two different rubrics. Similarly, Kim noted that while individual teachers could rank-order students comparably, actual rating agreement under identical conditions remained low to moderate. These observations underscore the challenges of attaining substantial inter-rater consistency in the inherently subjective domain of writing assessments.

The Average Measures ICC of 0.925, however, revealed excellent reliability when considering the mean scores across multiple raters. This high level of reliability implies that the diagnostic scale provides a consistent framework for assessing EFL writing proficiency when scores are averaged, minimizing individual rater biases and enhancing the tool's reliability as a group assessment instrument. Such a high ICC value indicates that the diagnostic scale functions effectively, offering reliable and consistent results when applied by multiple raters collectively. The high consistency of mean scores also aligns with previous findings on the reliability benefits of averaging scores in subjective assessments (Cicchetti & Sparrow, 1981; McGraw & Wong, 1996; Shrout & Fleiss, 1979).

While the Average Measures ICC demonstrates robust reliability for the scale, the moderate agreement observed in individual assessments underscores an inherent challenge in performance-based assessments—individual rater variability. This complexity indicates that, beyond aggregate reliability, attention must be directed toward understanding and mitigating the variability among raters. Despite comprehensive training, raters may still interpret criteria differently, reflecting the subjective nature of writing assessments. As Lukácsi

**The Journal of English Language Pedagogy and Practice**
Vol. 17, No.35, Autumn and Winter 2024
DOI: 10.71586/jal.2024.11181190893

(2021, p. 2) suggests, such variability in rater severity is inevitable and underscores a need for ongoing research into this "fact of life" to better understand and manage its impact on scoring consistency. Consequently, these findings highlight the importance of continuous rater training and development to mitigate such variability, while also recognizing the limitations of such efforts (Harsch & Martin, 2012; Knoch et al., 2021; Li, 2022; Lukácsi, 2021; Şahan & Razı, 2020).

This study's Pearson correlation analysis, addressing the fourth research question, examined the relationship between ratings assigned by various raters using the diagnostic scale and original scores from an instructor using IELTS Task 2 band descriptors. The strong to very strong positive correlations, ranging from 0.899 to 1.000, confirm the scale effectively measures the intended construct of EFL writing proficiency. Concurrent validity, as operationalized through these correlations, provides important evidence that the instrument reliably measures the same construct as the benchmark (Kim, 2010; Lin, 2024; Safari & Ahmadi, 2023).While all raters demonstrated strong positive correlations, reinforcing the scale's reliability, the perfect correlation ($r = 1.000$) observed for Rater 6 is noteworthy. This may reflect Rater 6's unique expertise as a professional certified IELTS examiner, which could influence scoring outcomes differently compared to other raters who, though experienced, may not have the same level of specialized training.

The findings underscore the indispensable need for comprehensive rater training, a point supported by research that highlights its importance in achieving consistency across various educational contexts and proficiency levels (Li, 2022; Şahan & Razı, 2020). Additionally, future studies could benefit from exploring the scale's performance across diverse writing samples using multifaceted Rasch modeling, which would help identify and adjust for potential sources of bias and variance in rater judgments. Despite these challenges, the scale's robust alignment with instructor judgments, supported by strong coefficient correlations, affirms its high concurrent validity. This suggests that the scale can effectively assess EFL writing proficiency, providing reliable and consistent evaluations even in the absence of an expert instructor. This capability aligns with prior research on diagnostic scale development, emphasizing the critical role of expert alignment in constructing practical assessment tools (Kim, 2019; Lukácsi, 2021; Safari & Ahmadi, 2023).

This study introduced an empirically designed rating scale aimed at providing targeted diagnostic feedback on EFL learners' academic writing skills. Built upon a rigorous foundation of empirical data collection and grounded in established

**The Journal of English Language Pedagogy and Practice**
Vol. 17, No.35, Autumn and Winter 2024
DOI: 10.71586/jal.2024.11181190893

theoretical models, the scale underwent extensive validation procedures, including content validity, inter-rater reliability, and concurrent validity, confirming its reliability and effectiveness. Unlike previous scales that often generalize across contexts, the descriptors developed here are tailored specifically to the EFL academic writing context, addressing critical elements such as content fulfillment, organizational knowledge, and language usage. By reflecting the specific challenges of EFL learners in academic settings, the diagnostic scale supports instructors in providing nuanced feedback that directly aligns with learners' unique developmental needs, filling a gap left by broader, less context-specific rating scales (e.g. Kim, 2019; Knoch, 2009; Turner & Upshur, 2002).

The implications of this study are substantial for EFL teaching and assessment. The diagnostic scale offers a detailed, empirically-based framework that enables educators to go beyond traditional scoring methods by identifying specific areas for improvement in academic writing (Cumming, 2001; Lumley, 2005; Weigle, 2002; Weir, 2005). This diagnostic orientation supports a learner-centered approach by offering concrete performance standards and meaningful, diagnostic feedback on essential academic writing skills. By identifying specific assessment elements tailored to EFL contexts, the scale allows educators to customize feedback to individual learners' needs, empowering students to self-assess their progress and take an active role in addressing their writing weaknesses (Brown & Harris, 2016). Furthermore, the structured descriptors could significantly enhance both large-scale and classroom-based assessments by providing standardized yet context-sensitive criteria that align with learners' academic requirements. This targeted feedback approach supports theories of adult learning, particularly self-directed (andragogy) and self-determined learning (heutagogy), which promote overall writing proficiency by helping learners clearly identify strengths and areas for improvement (Blaschke & Hase, 2019). The scale's empirical foundation and alignment with practical skill requirements make it a reliable resource for instructors aiming to implement consistent, diagnostic assessment in diverse educational settings.

While the results are promising, future research is encouraged to further explore the diagnostic scale's applicability across diverse EFL educational contexts and instructional settings. Conducting comparative studies could provide insights into the scale's adaptability and potential for broader applications, as well as inform refinements to its descriptors to enhance its usability and reliability. Additionally, qualitative research examining learner perceptions of diagnostic feedback from the scale would be beneficial in understanding its impact on learner

DOI: 10.71586/jal.2024.11181190893

motivation and engagement with the writing process. To accommodate varied learner needs and contexts, it is essential to explore how the scale might be adapted for different levels of proficiency or specific writing genres. Such adaptations could help tailor the diagnostic feedback to more precisely meet the developmental needs of learners across a spectrum of abilities and academic disciplines. Further, experimental studies comparing the effects of diagnostic feedback from this scale with traditional rubrics could be instrumental in measuring its direct impact on learner outcomes. These studies would provide empirical evidence to support the effectiveness of diagnostic approaches over more conventional methods, potentially influencing instructional practices in EFL settings.

In summary, this study demonstrates the potential of a diagnostic-oriented approach to writing assessment that is empirically grounded and contextually relevant to EFL learners. The findings underscore the need for continued research into developing diagnostic tools that support targeted feedback, enabling learners to identify specific areas for improvement in academic writing. Further investigation into diagnostic assessment tools is essential to bridge gaps in existing methodologies and to refine scales that provide precise, actionable insights, ultimately fostering learner autonomy and improving instructional practices in diverse EFL settings.

**Conflict of Interest:** None

## References

Alderson, J. C., Brunfaut, T., & Harding, L. (2015). Towards a theory of diagnosis in second and foreign language assessment: Insights from professional practice across diverse fields. *Applied Linguistics*, *36*(2), 236-260.

Baghestani, A. R., Ahmadi, F., Tanha, A., & Meshkat, M. (2019). Bayesian critical values for Lawshe's content validity ratio. *Measurement and Evaluation in counseling and Development*, *52*(1), 69-73.

Blaschke, L. M., & Hase, S. (2019). Heutagogy and digital media networks. *Pacific Journal of Technology Enhanced Learning*, *1*(1), 1-14.

Brown, G. T., & Harris, L. R. (Eds.). (2016). *Handbook of human and social conditions in assessment*. New York, NY: Routledge.

Cambridge Assessment English. (n.d.). *Placing students in the right exam*. Retrieved from https://www.cambridgeenglish.org.

DOI: 10.71586/jal.2024.11181190893

Cicchetti, D. V., & Sparrow, S. A. (1981). Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behavior. *American journal of mental deficiency*, *86*(2), 127-137.

Creswell, J. W. (2022). *A concise introduction to mixed methods research*. SAGE Publications.

Cumming, A. (2001). Learning to write in a second language: Two decades of research. *International journal of English studies*, *1*(2), 1-23.

Fraga-Vinas, L. (2022). Testing the reliability of two rubrics used in official English certificates for the assessment of writing. *Alicante Journal of English Studies/Revista Alicantina de Estudios Ingleses*, *36*, 85-109.

Fulcher, G. (1993). *The construction and validation of rating scales for oral tests in English as a foreign language* (Doctoral dissertation, University of Lancaster).

Grabe, W., & Kaplan, R. (1996). *Theory and practice of writing: An applied linguistic perspective*. Longman.

Greene, J. C., Caracelli, V. J., & Graham, W. F. (1989). Toward a conceptual framework for mixed-method evaluation designs. *Educational evaluation and policy analysis*, *11*(3), 255-274.

Hamp-Lyons, L. (1995). Rating nonnative writing: The trouble with holistic scoring. *TESOL Quarterly*, *29*(4), 759-762.

Hamp-Lyons, L. (2016). Farewell to holistic scoring. Part Two: Why build a house with only one brick? *Assessing Writing*, *100*(29), A1-A5.

Harding, L., Alderson, J. C., & Brunfaut, T. (2015). Diagnostic assessment of reading and listening in a second or foreign language: Elaborating on diagnostic principles. *Language Testing*, *32*(3), 317-336.

Harsch, C., & Martin, G. (2012). Adapting CEF-descriptors for rating purposes: Validation by a combined rater training and scale revision approach. *Assessing Writing*, *17*(4), 228-250.

Hayes, J. R. (1996). A new framework for understanding cognition and affect in writing. In C. M. Levy, & S. Ransdell (Eds.), *The science of writing: Theories, methods, individual differences, and applications* (pp. 1–27). Mahwah, NJ: Lawrence Erlbaum.

Jin, Y. (2023). The development and validation of the English writing enjoyment scale. *Perceptual and Motor Skills*, *130*(1), 555-575.

Kellogg, R. T., & Raulerson, B. A. (2007). Improving the writing skills of college students. *Psychonomic bulletin & review*, *14*, 237-242.

Khamboonruang, A. (2020). *Development and validation of a diagnostic rating scale for formative assessment in a Thai EFL university writing classroom: A mixed methods study* (Doctoral dissertation, The University of Melbourne). Retrieved from https://core.ac.uk/download/pdf/358463788.pdf

Khamboonruang, A. (2022). Building an Initial Validity Argument for Binary and Analytic Rating Scales for an EFL Classroom Writing Assessment: Evidence from Many-Facets Rasch Measurement. *rEFLections*, *29*(3), 675-699.

Kim, Y. H. (2010). *An argument-based validity inquiry into the empirically-derived descriptor-based diagnostic (EDD) assessment in ESL academic writing* (Doctoral dissertation, University of Toronto).

Kim, Y. H. (2019). Developing and validating empirically-derived diagnostic descriptors in ESL academic writing. *Journal of Asia TEFL*, *16*(3), 906-926.

Knoch, U. (2009). *Diagnostic writing assessment: The development and validation of a rating scale*. Frankfurt: Peter Lang.

Knoch, U., Deygers, B., & Khamboonruang, A. (2021). Revisiting rating scale development for rater-mediated language performance assessments: Modelling construct and contextual choices made by scale developers. *Language Testing*, *38*(4), 602-626.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, *15*(2), 155-163.

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology*, *28*(4), 563–575. https://doi.org/10.1111/j.1744-6570.1975.tb01393.x

Li, W. (2022). Scoring rubric reliability and internal validity in rater-mediated EFL writing assessment: Insights from many-facet Rasch measurement. *Reading and Writing*, *35*(10), 2409-2431.

Liljequist, D., Elfving, B., & Skavberg Roaldsen, K. (2019). Intraclass correlation–A discussion and demonstration of basic features. *PloS one*, *14*(7), e0219854.

Lin, J. (2024). Developing a reading proficiency scale for Chinese as a second language: a confirmatory factor analysis approach. *Language awareness*, *33*(3), 597-624.

Lu, Y., Han, Q., Fang, Z., & Shen, A. (2021). Development and Validation of a Diagnostic Rating Scale for EFL Writing in China. *International Journal of English Linguistics*, *11*(1).

Lumley, T. (2005). *Assessing second language writing: The raters' perspective.* Frankfurt: Peter Lang.

Lukácsi, Z. (2021). Developing a level-specific checklist for assessing EFL writing. *Language Testing*, *38*(1), 86–105. https://doi.org/10.1177%2F0265532220916703

Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing research*, *35*(6), 382-386.

Ma, X., Shi, X., Lu, C., & Li, R. (2022) Development and validation of a college English writing scale for classroom-based peer assessment. *Journal of Xi'an International Studies University*, *30*(1), 56–62. https://doi.org/10.16362/j.cnki.cn61-1457/h.2022.01.010

Mäkinen, K. (1995). Topic and Comment Development in EFL Compositions. *Studia Philologica Jyväskyläensia*, 35, 13-29.

Masuwai, A., Zulkifli, H., & Hamzah, M. I. (2024). Self-assessment for continuous professional development: The perspective of Islamic Education. *Heliyon*, *10*(19), 2308410. https://doi.org/10.1080/2331186X.2024.2308410

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological methods*, *1*(1), 30-46.

Mukminin, A., Habibi, A., Muhaimin, M., & Hidayat, M. (2023). Social media use for English writing (SMU-EW): Preservice English teachers. *Ampersand*, *10*, 100112.

North, B. (2003). *Scales for rating language performance: Descriptive models, formulation styles, and presentation formats*. Princeton, NJ: Educational Testing Service.

Park, H., & Yan, X. (2019). An investigation into rater performance with a holistic scale and a binary, analytic scale on an ESL writing placement test. *Papers in Language Testing and Assessment*, *8*(2), 34-64.

Perkins, K. (1983). On the use of composition scoring techniques, objective measures, and objective tests to evaluate ESL writing ability. *TESOL quarterly*, *17*(4), 651-671.

Polit, D. F., & Beck, C. T. (2006). The content validity index: are you sure you know what's being reported? Critique and recommendations. *Research in nursing & health*, *29*(5), 489-497.

Romero Jeldres, M., Díaz Costa, E., & Faouzi Nadim, T. (2023). A review of Lawshe's method for calculating content validity in the social sciences. *Frontiers in Education, 8*, 1271335.

Rubio, D. M., Berg-Weger, M., Tebb, S. S., Lee, E. S., & Rauch, S. (2003). Objectifying content validity: Conducting a content validity study in social work research. *Social Work Research, 27(2),* 94–104. https://doi.org/10.1093/swr/27.2.94

Rupp, A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. Guilford press.

Safari, F., & Ahmadi, A. (2023). Developing and evaluating an empirically-based diagnostic checklist for assessing second language integrated writing. *Journal of Second Language Writing*, *60*, 101007.

Şahan, Ö., & Razı, S. (2020). Do experience and text quality matter for raters' decision-making behaviors?. *Language Testing*, *37*(3), 311-332.

Sasaki, M., & Hirose, K. (1996). Explanatory variables for EFL students' expository writing. *Language Learning*, *46*(1), 137–174.

Shi, X., Ma, X., Du, W., & Gao, X. (2024). Diagnosing Chinese EFL learners' writing ability using polytomous cognitive diagnostic models. *Language Testing*, *41*(1), 109-134.

Shi, J., Mo, X., Sun, Z., 2012. Content validity index in scale development. J. Cent. S. Univ. 37 (2), 152–155. https://doi.org/10.3969/j.issn.1672-7347.2012.02.007

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, *86*(2), 420-428.

Terry, G., Hayfield, N., Clarke, V., & Braun, V. (2017). Thematic analysis. In C. Willig & W. S. Rogers (Eds.), *The SAGE Handbook of Qualitative Research in Psychology* (pp. 17-36). SAGE Publications Ltd. https://doi.org/10.4135/9781526405555.n2

Turner, C. E., & Upshur, J. A. (2002). Rating scales derived from student samples: Effects of the scale maker and the student sample on scale content and student scores. *Tesol Quarterly*, *36*(1), 49-70.

Upshur, J. A., & Turner, C. E. (1995). Constructing rating scales for second language tests. *ELT Journal*, *49*,3-12.

Wagner, M. (2015). *The centrality of cognitively diagnostic assessment for advancing secondary school ESL students' writing: A mixed methods study* (Doctoral dissertation, University of Toronto).

Wang, Y., & Xie, Q. (2022). Diagnosing EFL undergraduates' discourse competence in academic writing. *Assessing Writing, 53*, 100641.

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

Weir, C. J. (2005). Language testing and validation. *Hampshire: Palgrave McMillan, 10*, 9780230514577.

Weir, C. J., Chan, S. H. C., & Nakatsuhara, F. (2013). *Examining the criterion-related validity of the GEPT advanced reading and writing tests: Comparing GEPT with IELTS and real-life academic performance*. LTTC-GEPT Research Report, 1, 1–43.

Wilson, F. R., Pan, W., & Schumsky, D. A. (2012). Recalculation of the critical values for Lawshe's content validity ratio. *Measurement and Evaluation in Counseling and Development*, *45*(3), 197–210. https://doi.org/10.1177/0748175612440286

Zou, S., Yan, X., & Fan, J. (2024). Establishing analytic score profiles for large scale L2 writing assessment: The case of the CET-4 writing test. *Assessing Writing, 60*, 100826.

## Biodata

**Fatemeh Shoaei** is a PhD candidate in TEFL at Alborz Campus, University of Tehran, Iran. Her research interests include technology-enhanced language learning (TELL), second/foreign language acquisition, and language assessment.

**Sayyed Mohammad Alavi** is a professor at the Faculty of Foreign Languages and Literatures, University of Tehran, Iran. His research interests include applied linguistics, language teaching, and language testing and assessment. He has authored several books, including Data Analysis in Applied Linguistics, and has published research articles in Scopus-indexed journals such as System, Language Assessment Quarterly, and Journal of Multilingual and Multicultural Development.

**Hossein Karami** is an associate professor at the Faculty of Foreign Languages and Literatures, University of Tehran, Iran. His research interests focus on validity and fairness, particularly in the context of language testing. He is the author of Fairness Issues in Educational Assessment and has published extensively in journals such as Educational Research and Evaluation, RELC

**The Journal of English Language Pedagogy and Practice**
Vol. 17, No.35, Autumn and Winter 2024
DOI: 10.71586/jal.2024.11181190893

Journal, Psychological Test and Assessment Modeling, TESOL Journal, and Asia-Pacific Education Review.

## ساخت یک مقیاس تشخیصی-محور برای نوشتار دانشگاهی در آموزش زبان انگلیسی به عنوان زبان خارجی: یک رویکرد تجربی

با وجود علاقه روزافزون به ابزارهای ارزیابی تشخیصی در نوشتار زبان دوم، تحقیقات تجربی محدودی در خصوص توسعه این ابزارها برای زمینه‌های آموزش زبان انگلیسی به عنوان زبان خارجی (EFL) صورت گرفته است. این پژوهش، در پاسخ به این نیاز، به ساخت و رواسازی یک مقیاس تشخیصی-محور برای ارزیابی نوشتار دانشگاهی دانشجویان ایرانی می‌پردازد. با استفاده از روش تحقیق آمیخته، توصیف‌گرهای ضروری که منعکس‌کننده مهارت‌های اصلی نوشتار هستند، از طریق پروتکل تفکر با صدای بلند و بازخورد کارشناسان شناسایی شدند و سپس تجزیه و تحلیل‌های کمی برای اطمینان از پایایی و روایی انجام شد. یافته‌ها نشان می‌دهند که ۲۱ توصیف‌گر (descriptor) تجربیِ به‌دست‌آمده، جنبه‌های اساسی نوشتار دانشگاهی مانند تحقق محتوا، دانش سازماندهی و استفاده از زبان را پوشش می‌دهند و به مربیان امکان می‌دهند تا سطح مهارت زبان‌آموزان را با دقت بیشتری ارزیابی کنند. فرآیند رواسازی این مقیاس، شامل بررسی پایایی بین ارزیابان، روایی محتوا و روایی معیار-محور، اثربخشی آن را به‌عنوان ابزاری تشخیصی هم‌راستا با ارزیابی‌های کارشناسانه تأیید می‌کند. این ابزار به‌عنوان منبعی ارزشمند برای ارزیابی‌های گسترده و کاربردهای کلاسی به‌شمار می‌آید و با پشتیبانی از رویکردی دانشجو-محور، به دانشجویان کمک می‌کند تا با چالش‌های خاص نوشتاری خود روبرو شوند و آن‌ها را برطرف کنند.

**کلید واژه‌ها:** ساخت مقیاس، ارزیابی تشخیصی، نوشتار آکادمیک EFL، رویکرد تجربی، توصیف‌گر