ISSN 1799-2591 (print); ISSN 2053-0692 (online)
Theory and Practice in Language Studies, Vol. 16, No. 7, pp. 2544-2552, July 2026
DOI: https://doi.org/10.17507/tpls.1607.36

# Establishing Inter-Rater Reliability in CEFR-Based Textbook Evaluation: Evidence From a Fleiss' Kappa Study

Imed Sdiri
School of Languages, Literacies, and Translation, Universiti Sains Malaysia, Penang, Malaysia;
Ministry of Education, Directorate of Performance and Evaluation, Esa Town, Kingdom of Bahrain

Manjet Kaur Mehar Singh*
School of Languages, Literacies, and Translation, Universiti Sains Malaysia, Penang, Malaysia

***Abstract*—Reliability is a central concern in research on textbook evaluation, particularly when judgments depend on the subjective interpretation of pedagogical features. In the context of English as a Foreign Language (EFL) education, the Common European Framework of Reference for Languages (CEFR) has become the dominant benchmark for assessing curricular alignment. Yet, while numerous studies claim that textbooks are evaluated against CEFR guidelines, few report systematically on the reliability of such evaluations. This paper addresses this gap by examining inter-rater reliability in the evaluation of an EFL textbook marketed by a renowned international publisher as aligned with the CEFR A2 level. Three expert raters independently applied a validated 22-item evaluation instrument to 20 reading comprehension lessons. Agreement was measured using Fleiss' Kappa ($\kappa_F$), a statistic specifically designed for categorical data involving multiple raters. The results revealed a consistently high level of inter-rater reliability, with a pooled Fleiss' Kappa of 0.89. These findings confirm the robustness of the evaluation instrument and demonstrate that with rigorous rater training and independent coding, subjective bias in EFL textbook evaluation can be effectively minimized. The study contributes a methodological model that enhances the rigor of EFL textbook evaluation studies by advocating for evidence-based validation procedures.**

***Index Terms*—inter-rater reliability, Fleiss' Kappa, CEFR alignment, textbook evaluation, bias**

## I. Introduction

The evaluation of instructional materials has long been a central activity in the field of language education, serving as a critical mechanism for quality assurance and curriculum development (Sheldon, 1988; Tomlinson, 2012). In the context of English as a Foreign Language (EFL), textbooks are rarely just supplementary resources; they remain the backbone of teaching and learning, functioning as the 'de facto' curriculum in many classrooms (Tomlinson, 2012; Ponnusamy et al., 2021). In many educational systems around the world, the textbook dictates the scope and sequence of instruction, mediates the implementation of national policy, and defines the daily pedagogical reality for teachers and students. Consequently, ensuring the quality, suitability, and alignment of these materials with official policies and objectives is an essential task for researchers, policymakers, and practitioners alike (Mukundan & Ahour, 2010).

Over the past two decades, the landscape of EFL materials evaluation has been reshaped by the global adoption of the Common European Framework of Reference for Languages (CEFR). Originally developed by the Council of Europe (2001, 2020) and elaborated upon by North (2000, 2007, 2014, 2021), the CEFR has transcended its European origins to become the dominant international benchmark for language curriculum design and assessment (Byram & Parmenter, 2012; Kanchai, 2019; Fleckenstein et al., 2020; Khan et al., 2023). Its descriptive scheme, which defines proficiency through "can-do" statements across six vertical levels (A1 to C2), offers a shared metalanguage that promises transparency, comprehensiveness, and comparability (Council of Europe, 2001). In response to this global trend, many educational institutions have introduced CEFR-aligned curricula and textbooks, with the aim of benchmarking local student performance against common international guidelines.

However, the widespread adoption of the CEFR has introduced significant methodological challenges regarding the verification of alignment. While international publishers routinely label EFL textbooks with specific CEFR levels (e.g., A1, A2, B1), these claims are frequently accepted without rigorous empirical validation (Cambridge University Press, 2013). As some researchers have observed, the use of the CEFR often serves as a marketing tool rather than a verified pedagogical guarantee (Deygers et al., 2017; Ilc & Stopar, 2015). The fundamental difficulty lies in the nature of the framework itself. The CEFR descriptors are language-neutral (Alderson, 2005), which leads to under-specification and ambiguity (Kaur & Jian, 2022; Sufi & Idrus, 2021). Accordingly, significant researcher interpretation is required when

* Corresponding Author.

attempting to analyze specific texts or tasks in textbooks (Valax, 2011). For instance, evaluating whether a textbook or a test truly corresponds to a specific CEFR level has to involve complex expert judgments (Alderson, 2005), especially lexical load and cognitive demand as they play a key role in language learning (Laufer, 1989, 2021; Nation, 2006).

While explicit rubrics and checklists are often employed to guide these judgments, the evaluation process inevitably carries an element of subjectivity. Different evaluators, influenced by their own professional backgrounds and inner syllabi may interpret the same textbook, lesson, or task in divergent ways. This subjectivity raises a critical question that is often overlooked in materials development research: If experts cannot agree on the analysis of a task, can the textbook be considered validly aligned? Without transparent reporting on the degree of agreement among raters, known as inter-rater reliability (IRR), claims about alignment with the CEFR risk being undermined by questions of rigor and validity (Alderson, 2005).

Despite the importance of this issue, a review of the literature reveals a methodological gap. While numerous studies evaluate the content of textbooks against CEFR guidelines, few report systematically on the reliability of the evaluators themselves. For instance, Guerra et al. (2018) conducted a cross-national study in Turkey and Portugal using a 'Curriculum Evaluation Form' and a 'CEFR Checklist' to assess primary level coursebooks. While their study offered a comprehensive content analysis of A1 skills, it did not report statistical metrics regarding the consistency of the raters applying these checklists. Similarly, Demirel and Fakazli (2021) analyzed speaking and writing activities in EFL textbooks by categorizing them under CEFR A2 descriptors. Their methodology involved mapping tasks to 'criteria determined in the CEFR' - a process that inherently requires subjective interpretation of 'can-do' statements and design constraints (Council of Europe, 2001). Yet, the study relied on the researchers' categorization to determine the percentage of inclusion without providing inter-rater reliability coefficients to validate that these mappings were consistent. A comparable limitation is evident in Natova's (2021) study on estimating CEFR reading text complexity, where qualitative judgments regarding text purpose, structure, grammatical complexity, and knowledge demands were carried out by a single researcher and validated through learner performance rather than through independent expert agreement. In such cases, studies often rely on the 'consensus' of a panel or the assumption of expert competence without providing statistical evidence of agreement, leaving the robustness of their findings open to question.

To address this gap, the present study systematically examines inter-rater agreement in the evaluation of 20 reading comprehension lessons in an EFL textbook marketed by a renowned international publisher as aligned with the CEFR A2 level. By employing Fleiss' Kappa, a robust statistical measure designed for categorical judgments involving multiple raters, this research moves beyond descriptive analysis to quantify the consistency of expert judgment. In doing so, the paper not only strengthens the methodological transparency of CEFR alignment studies but also contributes to the broader discussion of how reliability can be established to minimize the double loss of learning gaps and financial waste caused by misaligned instructional materials.

## II. Literature Review

### *A. Reliability and the Challenge of Subjectivity in Educational Research*

Reliability is a fundamental pillar of rigorous educational research, serving as a prerequisite for validity (Bachman, 1990; McNamara, 2000). Creswell (2012) defines reliability as the stability and consistency of "individual scores" when a given research instrument is administered repetitively (p. 159). In quantitative research, this is often a matter of statistical replication; however, in qualitative and mixed-methods evaluations - such as the evaluation of textbook content - reliability becomes a matter of inter-subjectivity. It addresses whether different evaluators, viewing the same material through the lens of the same criteria, can arrive at consistent conclusions.

In the context of textbook evaluation, this is particularly fraught. Unlike checking a multiple-choice test key, evaluating whether a reading lesson aligns with a given CEFR descriptor or guideline requires interpretation. If Rater A perceives a lesson or a task as aligned with the CEFR guidelines, while Rater B perceives it as misaligned, the validity of the evaluation is compromised. Without established reliability, the resulting data reflects the idiosyncrasies of the raters rather than the intrinsic qualities of the textbook. Therefore, establishing high inter-rater reliability is not merely a statistical exercise but a necessary safeguard to ensure that claims about textbook quality are defensible and independent of any form of individual observer bias.

### *B. Mitigating Subjectivity: Independent Expert Coding and Triangulation*

To mitigate the subjectivity inherent in expert judgment and qualitative research, this study adopts a rigorous multi-rater protocol rooted in the principle of triangulation. In the context of materials evaluation, triangulation may include the deployment of multiple observers to examine the same object of study, thereby reducing the potential for idiosyncratic bias associated with a single researcher (Denzin, 1978; Patton, 2002).

Rather than relying on a solitary evaluation, which may be influenced by personal pedagogical preferences, this study employs independent parallel coding. In this approach, a panel of experts assesses the materials in isolation using a standardized instrument. By ensuring that raters work independently prior to any statistical analysis, the methodology transforms subjective individual opinions into objective, quantifiable data points (Krippendorff, 2018). This process safeguards the validity of the evaluation, as high levels of consistent agreement among independent raters across the

various units of analysis provide strong empirical evidence that the textbook's alignment (or misalignment) is an observable property and a confirmed fact, rather than an artifact of rater bias.

### *C. Threats to Validity: The 'Human Instrument' and Cognitive Bias*

While expert judgment is essential for evaluating complex pedagogical constructs, it is susceptible to various forms of research bias that can severely undermine the reliability and validity of the findings. In qualitative research, the researcher or rater acts as the human instrument for data collection. This makes the process vulnerable to systematic errors in cognition and social interaction (Creswell, 2012; Podsakoff et al., 2003).

One of the most pervasive threats is social desirability bias. In the context of inter-rater reliability studies, this occurs when raters unconsciously align their judgments with what they perceive to be the expectations of the lead researcher or the consensus of the group, rather than their own independent assessment (Nederhof, 1985). If raters discuss their evaluations during the coding process, they may suppress dissenting views to maintain group harmony, leading to artificially inflated agreement scores that do not reflect true reliability.

Furthermore, specific cognitive biases pose significant risks in textbook evaluation:

- Confirmation Bias: Researchers may preferentially notice evidence that supports their pre-existing beliefs (Nickerson, 1998; Peters, 2022). For instance, if a researcher believes that international publishers always produce high-quality work, they may overlook misalignments in the text.
- The Halo Effect: This cognitive bias makes an initial positive impression shape later judgments and decisions across other unrelated attributes (Schuldt et al., 2012; Schouteten et al., 2019). For instance, a rater's positive impression of one aspect of the textbook (e.g., high-quality graphic design or layout) may unduly influence their judgment of unrelated factors, such as pedagogical rigor or CEFR alignment.
- Prestige Bias: Some researchers tend to apply less scrutiny to scholarly work when it comes from well-known individuals or prestigious institutions (Cranford, 2020). In the context of the current study, the reputation of the publisher or the CEFR label itself can increase acceptance and trust, causing raters to assume alignment exists simply because it is claimed on the cover.

To minimize these threats, methodological rigor is required. This includes the use of independent coding protocols (where raters work in isolation to prevent social influence) and statistical validation to detect whether the observed agreement is genuine or merely an artifact of bias.

### *D. Inter-Rater Agreement and Reliability: Methodological Considerations and Applications*

Quantifying inter-rater agreement (IRA) or inter-rater reliability (IRR) is crucial for validating assessment procedures across diverse scientific fields, particularly wherever subjective judgment is involved, and a definitive gold standard is unavailable (Gwet, 2014; Benchoufi et al., 2020; Nelson & Pepe, 2000). IRA measures the degree of concordance among independent raters classifying subjects or objects (Krippendorff, 2016; Klein, 2018). This quantification is essential in areas such as medical diagnoses, psychiatric evaluations, interpretation of diagnostic images, surgical reviews, and evaluations in educational contexts (Benchoufi et al., 2020; Moons & Vandervieren, 2025) as high agreement suggests clarity and objectivity in the underlying guidelines used for classification.

### *E. Foundational Kappa Statistics*

The most common statistical measure used to quantify agreement beyond chance is the kappa coefficient (Cohen, 1960; Fleiss, 1971; Klein, 2018). For two raters classifying outcomes on a nominal scale, Cohen's kappa ($K_C$) compares the observed proportion of agreement ($P_o$) against the proportion of agreement expected by chance ($P_e$), normalizing the difference by the maximum possible improvement over chance ($1 - P_e$) (Cohen, 1960; Fleiss et al., 2003; Klein, 2018). $K_C$ accounts for the possibility of differing marginal classification probabilities between the two raters (Cohen, 1960).

When more than two raters are involved, researchers typically rely on the Fleiss kappa ($K_F$) which is a popular generalization derived from the concept of pairwise agreement (Fleiss, 1971; Klein, 2018; Benchoufi et al., 2020). As a statistical test, $K_F$ is used to verify the trustworthiness of collected data (Fleiss, 1971) and accordingly monitor the biases mentioned above. It measures the level of inter-rater agreement among different raters (Hassan et al., 2019; McHugh, 2012). While earlier measures like Cohen's Kappa (Cohen, 1960) are suitable for dyadic agreement, Fleiss' Kappa is preferred for categorical data with multiple raters (Gwet, 2014).

## III. Methodology

### *A. Research Context and Corpus*

This study evaluated an EFL textbook marketed by a renowned international publisher as aligned with the CEFR A2 level. Many educational systems around the world rely heavily on imported CEFR-aligned textbooks to deliver the curriculum, as they believe that such materials are of high quality. This assumption requires in-depth validation to make sure that such claims are based on facts and not mere labels on EFL textbook covers. The analysis focused specifically on the 20 reading comprehension lessons contained within the textbook. These comprise a total of 114 associated tasks representing the core input for reading comprehension proficiency development for A2 learners.

### *B. Evaluation Instrument*

To assess alignment between the reading comprehension lessons and the CEFR guidelines, a validated evaluation form was developed based on different parameters and criteria included in different CEFR-related documents that were officially published by the Council of Europe. These included the 2001 and 2020 versions of the CEFR and the Guide for Users (Bailly, 2003) which was published by the Council of Europe in 2001. The form comprised 22 criteria organized under 12 main parameters reflecting critical aspects of CEFR-based material design. These parameters addressed: (1) the approach, (2) context of language use, (3) conditions and constraints, (4) learners' mental context, (5) communication themes, (6) texts, (7) learners' competence, (8) learners' general needs and wants, (9) desirable and practical media to present the material, (10) grouping and sequencing, (11) task range, and (12) effective task design. Prior to data collection, the instrument underwent content validation by a panel of university scholars and EFL experts to ensure the criteria were operationally clear and distinct.

To ensure objective classification suitable for inter-rater reliability analysis, the instrument employed a strict categorical coding scheme. For the majority of the criteria (20 items), raters utilized a binary scale (1 = Criterion Met/Yes; 2 = Criterion Not Met/No) to assess the presence or absence of specific CEFR design features. Two criteria utilized multinomial nominal scales to categorize specific lesson attributes: Criterion 2 required raters to identify the dominant CEFR Domain (coded as 1=Personal, 2=Public, 3=Occupational, 4=Educational), and Criterion 17 required the classification of Task Type (coded as 1=Pedagogic, 2=Authentic). This structure ensured that all generated data was nominal, fulfilling the statistical assumptions required for the calculation of Fleiss' Kappa.

### *C. Raters and Training Protocol*

The evaluation was conducted by a panel of three raters with long experience in EFL curriculum development, textbook evaluation, and EFL teaching and learning. To minimize the risk of rater drift and ensure a unified understanding of the evaluation criteria, prior to the main study, a comprehensive training workshop was conducted by a Canadian EFL expert who has worked with renowned international publishers like Pearson and York Press.

The workshop focused on deepening understanding and standardizing interpretation of the different parameters and criteria included in the evaluation form. Besides, a bank of key reference documents was shared with the raters. These included the Target Textbook, the 2001 and 2020 versions of the CEFR, and the Guide for Users. Special emphasis was particularly given to the need to stick to the evaluation guidelines as introduced in the training workshop and as reflected in the evaluation form. This training phase was essential to mitigate cognitive biases and ensure that all raters approached the material with a shared theoretical framework.

### *D. Quality Assurance: Piloting and Calibration*

To further enhance reliability, a pilot study was conducted using a representative sample unit from the textbook. The three raters independently evaluated this sample unit, after which their ratings were analyzed and compared. This pilot phase led to the refinement of a few criteria for clarity. It also proved the potential of the evaluation form to generate the required data.

Following the pilot, the main data collection employed a rigorous, reiterative process. Raters performed their tasks independently to avoid any form of influence or bias. After each lesson was analyzed, the main researcher reviewed the responses and measured inter-rater reliability.

### *E. Statistical Analysis*

Fleiss' Kappa ($\kappa_F$) was calculated to quantitatively measure the level of agreement beyond chance across the three evaluators. This statistic was chosen over percent agreement because it corrects for chance and, accordingly, provides a more conservative and robust index of reliability. The interpretation of Kappa values followed the standard guidelines that several researchers have followed (Landis & Koch, 1977; Hassan et al., 2019; Laerd Statistics, 2019; Albakkosh, 2024). Values < 0.00 indicate poor agreement, 0.01–0.20 slight agreement, 0.21–0.40 fair agreement, 0.41–0.60 moderate agreement, 0.61–0.80 substantial agreement, and 0.81–1.00 almost perfect agreement. All statistical analyses were conducted using SPSS software to ensure computational accuracy.

## IV. RESULTS

### *A. Inter-Rater Reliability Analysis*

The primary objective of the statistical analysis was to determine the extent to which the three expert raters applied the evaluation instrument consistently when assessing the CEFR alignment of the 20 reading comprehension lessons. Inter-rater reliability was examined using Fleiss' Kappa **($\kappa_F$)**, a chance-corrected coefficient specifically designed to measure agreement among multiple raters assigning categorical judgments. Unlike simple percentage agreement, Fleiss' Kappa explicitly adjusts for the level of agreement that would be expected to occur by chance alone, thereby providing a more conservative and statistically rigorous estimate of reliability.

Fleiss' Kappa coefficients were calculated separately for each lesson to capture lesson-specific variations in agreement and to identify any localized sources of rater divergence. In addition, a pooled Fleiss' Kappa was computed across all lessons to provide an overall estimate of agreement that accounts simultaneously for observed consensus and chance

agreement across the entire corpus. Table 1 presents the detailed Fleiss' Kappa coefficients for each lesson, alongside the pooled measure for the entire corpus.

TABLE 1
FLEISS' KAPPA IN READING COMPREHENSION LESSONS

| Unit and Lesson Number | Fleiss' Kappa |
|---|---|
| Unit 1, Lesson 2 | 0.83 |
| Unit 1, Lesson 5 | 0.79 |
| Unit 2, Lesson 2 | 0.85 |
| Unit 2, Lesson 5 | 0.86 |
| Unit 2, Lesson 11 | 0.92 |
| Unit 3, Lesson 2 | 0.80 |
| Unit 3, Lesson 5 | 0.81 |
| Unit 4, Lesson 2 | 0.87 |
| Unit 4, Lesson 5 | 0.87 |
| Unit 4, Lesson 11 | 0.88 |
| Unit 5, Lesson 2 | 0.96 |
| Unit 5, Lesson 5 | 0.94 |
| Unit 6, Lesson 2 | 0.96 |
| Unit 6, Lesson 5 | 0.92 |
| Unit 6, Lesson 11 | 0.91 |
| Unit 7, Lesson 2 | 0.91 |
| Unit 7, Lesson 5 | 0.95 |
| Unit 8, Lesson 2 | 0.89 |
| Unit 8, Lesson 5 | 0.96 |
| Unit 8, Lesson 11 | 0.87 |
| Pooled Fleiss' Kappa | 0.89 |

*B. Analysis of Agreement Patterns*

The data reveals a robust consensus among the three raters, with a Pooled Fleiss' Kappa of 0.89. According to the established benchmarks, this indicates an 'almost perfect' level of chance-corrected agreement overall. This high pooled value confirms that the evaluation form was operationally clear and that the raters applied the CEFR parameters and criteria with a high degree of consistency across the majority of the corpus. A granular analysis of the lesson-specific scores reveals three distinct patterns of inter-rater reliability:

*(a). The Perfect Case ($\kappa_F$ = 1.00)*

This level of agreement was not recorded in any lesson. This finding is neither unexpected nor problematic. In applied research, particularly when expert judgment is required to evaluate multidimensional constructs such as CEFR alignment, perfect agreement is rare. The absence of $\kappa_F$ = 1.00 reflects the inherently interpretive nature of textbook evaluation rather than deficiencies in the evaluation framework or rater expertise.

*(b). Near-Perfect Consensus ($\kappa_F$ > 0.81)*

The vast majority of the corpus (19 out of 20 lessons) achieved scores in the 'almost perfect' range. $\kappa_F$ scores varied from $\kappa_F$ = 0.80 to $\kappa_F$ = 0.96. This suggests that the target lessons in the textbook possessed unambiguous design features—such as explicit domain classification, attractive design, task variety, task sequencing, etc.—that left little room for deviations and inter-rater agreement. In these instances, the textbook's alignment (or misalignment) with the CEFR was objectively observable to all raters. The high $\kappa_F$ values therefore indicate not merely surface-level consensus, but systematic agreement grounded in shared interpretation of the evaluation framework.

*(c). Substantial Agreement ($\kappa_F$ =0.61–0.80)*

This level of agreement was recorded in Unit 1, Lesson 5 ($\kappa_F$ = 0.79). While still indicating 'Substantial' agreement, this slight disagreement suggests that this specific lesson may have contained elements that are not that evident for interpretive judgment. However, this variation was minimal and did not compromise the overall reliability of the evaluation. Crucially, because Fleiss' Kappa penalizes agreement that could occur randomly, the $\kappa_F$ value of 0.79 confirms that the observed divergence was modest and systematic rather than random.

Overall, the predominance of scores above 0.80 demonstrates that the independent multi-rater protocol and the accompanying rater training were highly effective in standardizing expert judgment. The data confirms that while isolated instances of ambiguity exist within the textbook, the evaluation framework itself is statistically reliable and robust for assessing CEFR alignment.

*C. Summary of Chance-Corrected Reliability Findings*

The predominance of high Fleiss' Kappa values confirms that the observed inter-rater agreement was not an artifact of chance but a product of consistent and principled application of the evaluation instrument. The standardized evaluation

protocol, combined with structured rater training, appears to have successfully minimized random variation and aligned expert interpretations of CEFR-based parameters and criteria.

These results establish that the evaluation framework succeeded in yielding statistically reliable, chance-corrected judgments. This reinforces the methodological credibility of the study and provides a robust foundation for subsequent analyses to measure the extent to which the various reading comprehension lessons included in the Target Textbook are aligned with the CEFR guidelines regarding textbook design.

## V. Discussion

### *A. High Reliability Through Rigorous Methodology*

The overall pooled Fleiss' Kappa of 0.89 represents a significant finding in the field of EFL materials evaluation. In qualitative research, where judgments rely heavily on the interpretation of pedagogical constructs, achieving such 'almost perfect' agreement is promising. This result suggests that high reliability is not merely a product of rater expertise but the outcome of a deliberate, structured quality assurance protocol.

The robust consensus observed here is directly attributable to the multi-stage quality assurance process. The success in designing a well-structured and clear evaluation form that managed to include core CEFR design parameters and criteria gave the raters strong support to understand what is required. Besides, the initial training workshop served to align the raters' internal benchmarks with the external guidelines of the CEFR, effectively systematizing the 'human instrument' before data collection began. Furthermore, the pilot study acted as a critical stress test by allowing for the identification and resolution of ambiguities in the instrument prior to the main analysis.

This quality assurance process was further enhanced by constant monitoring and reflexivity. Rather than treating the evaluation as a static event, the study engaged in an iterative process of reviewing preliminary ratings and calculating interim reliability metrics. This allowed for the detection of rater drift—where evaluators unconsciously shift their standards over time. By combining independent evaluation with reflective oversight, the methodology successfully balanced the need for individual judgment with the requirement for acceptable levels of inter-rater reliability. While high agreement does not strictly or fully eliminate the prospect of shared systematic bias among raters, the rigorous independent coding protocol ensures that the findings represent a consistent, replicable professional consensus rather than idiosyncratic opinion.

### *B. Implications for Textbook Evaluation Research*

The high reliability scores validate the 22-item instrument used in this study. They demonstrate that even complex, high-inference CEFR constructs—such as "task sequencing" and "relevance to learners' competence"—can be evaluated consistently. However, the findings imply that the instrument alone is insufficient; it must be paired with a rigorous calibration protocol where raters are offered the opportunity to deepen their understanding of the evaluation form and its various items.

In addition to validating the instrument used in the study, the findings have profound implications for future research. Studies that rely on "checklist" evaluations without measuring inter-rater reliability, resorting to rater training, or piloting the study before actual implementation may be masking significant subjective variance. The findings the present study has generated suggest that reliability is not an inherent property of an evaluation tool, but rather an outcome of the interaction between a well-designed instrument and a well-trained evaluator. Future studies should therefore prioritize reflexivity as stressed in qualitative research (Riazi, 2016), explicitly documenting how raters' understanding of the criteria was enhanced. This shift from single unchecked evaluations to transparent methodological reporting is essential for establishing dependability in applied linguistics research, and other research fields, as advocated in relevant methodological frameworks.

### *C. Practical Applications for Policy and Governance*

For policymakers and publishers, these results offer a new standard for accountability. The demonstration that textbook alignment can be measured with statistical precision challenges the prevailing reliance on impressionistic reviews. For Ministries of Education, this provides a mechanism for evidence-based governance. Decisions regarding curriculum reform and the procurement of international textbooks involve substantial public investment (Setyono & Widodo, 2019). When multiple experienced raters independently agree that a lesson meets (or fails to meet) CEFR guidelines, the resulting data provides a defensible basis for high-stakes resource allocation. It minimizes the financial and educational waste associated with adopting misaligned materials. For EFL textbook publishers, including the most renowned of them, this study highlights the necessity of subjecting their "CEFR-aligned" claims to external, statistical validation before bringing products to the international market. This required shift in EFL textbook production and marketing is essential to ensure alignment with CEFR guidelines and relevance to students' needs.

## VI. Conclusion

This study addressed the critical challenge of subjectivity in materials evaluation by examining inter-rater agreement in the evaluation of an EFL textbook marketed by a renowned international publisher as aligned with the CEFR A2 level.

By applying Fleiss' Kappa to the judgments of three independent raters, the analysis revealed a pooled agreement coefficient of 0.89. This 'almost perfect' reliability score serves as empirical validation not only of the textbook's alignment features but, more importantly, of the evaluation instrument, the rigorous calibration process, and the high level of chance-corrected inter-rater reliability. It demonstrates that expert judgment in qualitative language research, when properly trained and statistically monitored, can provide a stable and objective basis for assessing EFL teaching and learning materials.

From a methodological perspective, this research bridges the gap between qualitative interpretation and quantitative rigor. By moving beyond simple percentage agreement to report chance-corrected metrics, the study provides a transparent, granular picture of expert consensus. It confirms that the 'human instrument' in qualitative research need not be a source of error; rather, through rigorous training, independent coding, and reiterative reflexivity, it can become a reliable tool for high-stakes evaluation. The study thus offers a replicable blueprint for future researchers seeking to validate CEFR alignment claims with statistical precision.

Beyond the academy, the implications for language policy and governance are profound. In centralized education systems where EFL textbooks are used as main tools to implement the intended curriculum, the cost of adopting misaligned materials is high - measured not only in financial terms but in learning loss and human capital stagnation. This study argues that reliance on publisher claims or impressionistic reviews is no longer tenable. Instead, education ministries and language institutions providing EFL services must institutionalize evidence-based validation procedures. If the CEFR is to serve as a meaningful benchmark for global language education, the verification of alignment must shift from a marketing label to a rigorous, data-driven standard of practice.

## ACKNOWLEDGEMENTS

The authors would like to express their sincere gratitude to the expert raters who contributed their time and expertise to the validation phase of this study using a rigorous quality assurance protocol.

## REFERENCES

[1] Albakkosh, I. (2024). Using Fleiss' Kappa Coefficient to Measure the Intra and Inter-Rater Reliability of Three AI Software Programs in the Assessment of EFL Learners' Story Writing. *International Journal of Educational Sciences and Arts*, *3*(1), 69–96. https://doi.org/10.59992/ijesa.2023.v3n1p4

[2] Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. Continuum.

[3] Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.

[4] Bailly, S. (Ed.). (2003). *Common European Framework of Reference for Languages: Learning, teaching, assessment; A guide for users*. Council of Europe.

[5] Benchoufi, M., Matzner-Lober, E., Molinari, N., Jannot, A. S., & Soyer, P. (2020). Interobserver agreement issues in radiology. *Diagnostic and interventional imaging*, *101*(10), 639-641. https://doi.org/10.1016/j.diii.2020.09.001

[6] Byram, M., & Parmenter, L. (Eds.). (2012). *The Common European Framework of Reference: The globalisation of language education policy*. Multilingual Matters.

[7] Cambridge University Press. (2013). *Introductory guide to the Common European Framework of Reference (CEFR) for English language teachers*. Cambridge University Press.

[8] Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46.

[9] Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.

[10] Council of Europe. (2020). *Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion volume*. Council of Europe Publishing.

[11] Cranford, S. (2020). The Pursued, the Pursuing, and Unconscious Prestige Bias. *Matter*, *2*(5), 1065-1067.

[12] Creswell, J. W. (2012). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (4th ed.). Pearson.

[13] Demirel, E., & Fakazli, Ö. (2021). A comparison of the speaking and writing activities in EFL coursebooks with the CEFR. *International Journal of Curriculum and Instruction*, *13*(1), 168–189.

[14] Deygers, B., Zeidler, B., Vilcu, D., & Carlsen, C. H. (2017). One Framework to Unite Them All? Use of the CEFR in European University Entrance Policies. *Language Assessment Quarterly*, *15*(1), 3–15. https://doi.org/10.1080/15434303.2016.1261350

[15] Fleckenstein, J., Keller, S., Krüger, M., Tannenbaum, R. J., & Köller, O. (2020). Linking TOEFL iBT® writing rubrics to CEFR levels: Cut scores and validity evidence from a standard setting study. *Assessing Writing*, *43*, 100420. https://doi.org/10.1016/j.asw.2019.100420

[16] Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378–382.

[17] Fleiss, J. L., Levin, B., & Paik, M. C. (2003). Statistical inference for a single proportion. *Statistical Methods for Rates and Proportions*, *3*, 64-79. https://doi.org/10.1002/0471445428.ch2

[18] Guerra, J., Gonçalves, S., Fisne, F. N., & Gungor, M. (2018). The CEFR in primary English classrooms: A snapshot from Turkey and Portugal. *Eurasian Journal of Educational Research*, *76*, 123–144.

[19] Gwet, K. L. (2014). *Handbook of inter-rater reliability* (4th ed.). Advanced Analytics.

[20] Hassan, E., Miller, P., & Jiang, D. (2019). Inter-rater reliability in qualitative research. *Journal of Research Practice*, *15*(2), 110–125.

[21] Ilc, G., & Stopar, A. (2015). Validating the Slovenian national alignment to CEFR: The case of the B2 reading comprehension examination in English. *Language Testing*, *32*(4), 443–462.

[22] Kaur, P., & Jian, M. Z. (2022). The CEFR-aligned curriculum: Perspectives of Malaysian teachers. *Asian Journal of Research in Education and Social Sciences*, *4*(1), 138–145.
[23] Kanchai, T. (2019). Thai EFL university lecturers' viewpoints towards impacts of the CEFR on their English language curricula and teaching practice. *NIDA Journal of Language and Communication*, *24*(35), 23–47. Retrieved January 16, 2026, from https://so04.tci-thaijo.org/index.php/NJLC/article/download/202408/141228
[24] Khan, A., David, A. R., Ahmad, A. H., Ali, A., & Lah, S. C. (2023). Initial Insights into CEFR Adoption at a Language Faculty of a Public University in Malaysia. *PASAA*, *67*(1), 330-360. Retrieved January 16, 2026, from https://digital.car.chula.ac.th/cgi/viewcontent.cgi?article=1795&context=pasaa
[25] Klein, D. (2018). Implementing a General Framework for Assessing Interrater Agreement in Stata. *The Stata Journal: Promoting Communications on Statistics and Stata*, *18*(4), 871-901. https://doi.org/10.1177/1536867X1801800408
[26] Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174.
[27] Laerd Statistics. (2019). *Fleiss' Kappa in SPSS Statistics*. Retrieved January 16, 2026, from https://statistics.laerd.com/
[28] Laufer, B. (1989). What percentage of text lexis is essential for comprehension? In C. Lauren & M. Nordman (Eds.), *Special language: From humans thinking to thinking machines* (pp. 316–323). Multilingual Matters.
[29] Laufer, B. (2021). Lexical thresholds and alleged threats to validity: A storm in a teacup. *Reading in a Foreign Language*, *33*(2), 238–246.
[30] McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, *22*(3), 276–282.
[31] McNamara, T. (2000). *Language testing*. Oxford University Press.
[32] Moons, F., & Vandervieren, E. (2025). Measuring agreement among several raters classifying subjects into one or more (hierarchical) categories: A generalization of Fleiss' kappa. *Behavior Research Methods*, *57*(10), 287. https://doi.org/10.3758/s13428-025-02746-8
[33] Mukundan, J., & Ahour, T. (2010). A review of textbook evaluation checklists across four decades (1970-2008). *Porta Linguarum*, *13*, 336–352.
[34] Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review*, *63*(1), 59–82.
[35] Natova, I. (2021). Estimating CEFR reading comprehension text complexity. *The Language Learning Journal*, *49*(6), 699–710.
[36] Nederhof, A. J. (1985). Methods of coping with social desirability bias: A review. *European Journal of Social Psychology*, *15*(3), 263–280.
[37] Nelson, J. C., & Pepe, M. S. (2000). Statistical description of interrater variability in ordinal ratings. *Statistical methods in medical research*, *9*(5), 475-496.
[38] Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175–220.
[39] North, B. (2000). *The development of a common framework scale of language proficiency*. Peter Lang.
[40] North, B. (2007). The CEFR illustrative descriptor scales. *The Modern Language Journal*, *91*(4), 656-659.
[41] North, B. (2014). *The CEFR in practice* (Vol. 4). Cambridge University Press.
[42] North, B. (2021). The CEFR companion volume—What's new and what might it imply for teaching/learning and for assessment. *CEFR Journal-research and practice*, *4*, 5-24.
[43] Patton, M. Q. (2002). *Qualitative research & evaluation methods* (3rd ed.). Sage Publications.
[44] Peters, U. (2022). What is the function of confirmation bias? *Erkenntnis*, *87*(3), 1351-1376. https://doi.org/10.1007/s10670-020-00252-1
[45] Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*(5), 879–903.
[46] Ponnusamy, N. K., Sandaran, S. C., & Gunasegaran, I. (2021). Evaluation of Year 6 KSSR English (SK) textbook: Teachers' perspectives. *LSP International Journal*, *8*(1), 67–80.
[47] Riazi, A. M. (2016). *The Routledge encyclopedia of research methods in applied linguistics*. Routledge.
[48] Schouteten, J. J., Gellynck, X., & Slabbinck, H. (2019). Influence of organic labels on consumer's flavor perception and emotional profiling: Comparison between a central location test and home-use-test. *Food Research International*, *116*, 1000-1009. https://doi.org/10.1016/j.foodres.2018.09.038
[49] Schuldt, J. P., Muller, D., & Schwarz, N. (2012). The "fair trade" effect: Health halos from social ethics claims. *Social Psychological and Personality Science*, *3*(5), 581-589. https://doi.org/10.1177/1948550611431
[50] Setyono, B., & Widodo, H. P. (2019). The representation of multicultural values in the Indonesian Ministry of Education and Culture-endorsed EFL textbook: A critical discourse analysis. *Intercultural Education*, *30*(4), 383–397. https://doi.org/10.1080/14675986.2019.1658761
[51] Sheldon, L. E. (1988). Evaluating ELT textbooks and materials. *ELT Journal*, *42*(4), 237–246.
[52] Sufi, M. K. A., & Idrus, F. (2021). A preliminary study on localising the CEFR written production descriptor to Malaysian higher education context. *Asian Journal of Research in Education and Social Sciences*, *3*(2), 1–15.
[53] Tomlinson, B. (2012). Materials development for language learning and teaching. *Language Teaching*, *45*(2), 143–179. https://doi.org/10.1017/S0261444811000528
[54] Valax, P. (2011). *The Common European Framework of Reference for Languages: A critical analysis of its impact on a sample of English language teaching material* [Doctoral dissertation, University of Waikato].

**Imed Sdiri**, originally from Tunisia, is a senior specialist, university lecturer, and EFL textbook author and reviewer with over 20 years of experience in Bahrain, including advisory service at the Education Minister's Office. He is the founder and director of Bookiverse.online, a digital platform offering CEFR-aligned reading programs. His research focuses on applied linguistics, CEFR applications, readability studies, literacy, language planning, quality education, and human capital development.

**Manjet Kaur Mehar Singh** is an Associate Professor at the School of Languages, Literacies and Translation, Universiti Sains Malaysia. Her interests include sociolinguistics, language teaching and learning, academic literacy(ies), discourse, and multilingualism. She also manages the International Journal of Language, Literacy and Translation (IJoLLT) as the Chief Editor. Manjet Kaur is listed by Britishpedia (4th edition, 2022) as one of the "Successful People in Malaysia' in the field of Education.