Variations and Methodological Components in CEFR-Aligned Language Tests: A Systematic Review

Nik Aloesnita binti Nik Mohd Alwi*
aloesnita@umpsa.edu.my
Centre For Modern Languages
Universiti Malaysia Pahang Al-Sultan Abdullah, Malaysia

Wan Alisa Hanis binti Wan Abdul Halim
aleesahanes@gmail.com
Centre For Modern Languages
Universiti Malaysia Pahang Al-Sultan Abdullah, Malaysia

Corresponding author*

CITE THIS ARTICLE:

ABSTRACT
The Common European Framework of Reference for Languages (CEFR) has provided a flexible framework for constructing and implementing language tests. Its flexibility has led to the development of various CEFR-aligned language tests that attempted to conform to its framework. The variability in test purpose, quality, and difficulty has necessitated the use of different methodological decisions when conducting studies on CEFR-aligned language tests. The current study employed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines to conduct a present systematic review. A total of 31 journal articles on CEFR-aligned language tests were used to explore the methods employed in the articles. The review identified six prominent discoveries: the study method, data analysis, participant characteristics, rater participation, number of essays, and writing genres. The current study revealed the prevalence of quantitative methods, the focus on specific participant characteristics, the inclusion of many raters and essays, and the inclination towards two dominant writing genres in studies on CEFR-aligned language tests. By discovering the shared methodological components adopted in current CEFR-aligned language studies, this study guides researchers and practitioners in enhancing the validity, reliability, and generalisability of studies on CEFR-aligned language tests in various contexts through successful replicability.

INTRODUCTION

The Common European Framework of Reference for Languages, commonly known as CEFR, is a fundamental framework language educators utilise for conceptualising language learning. The framework serves as a guiding document in developing language learning curricula, teaching materials, and assessments (Nagai, 2020). Although the CEFR has been extensively adopted as a reference in the development of assessments, it is crucial to acknowledge that it was not initially intended to be used as a tool for developing standardised tests. Mendoza and Knoch (2018) highlight that the initial version of the CEFR lacked the necessary specificity to facilitate assessments, as its scales were too broad and failed to provide insights into task-specific competencies, making its implementation on language tests challenging. Nevertheless, there has been a notable trend in Europe, where many language tests have either aligned their scoring systems with the CEFR or undergone redevelopment with the CEFR descriptors in mind (Deygers et al., 2017). In this paper, these tests are referred to as CEFR-aligned language tests, indicating that they have been designed and implemented in accordance with the guidelines and principles outlined by the CEFR.

CEFR affords flexibility in the construction and administration of language tests. This alignment has led to a variety of task quality, nature, and difficulty in CEFR-aligned language tests developed by test providers and examination boards. According to Mat Yusoff et al. (2022), the flexibility accommodates the different language needs and cultural differences worldwide, allowing the use of CEFR throughout the world. Despite studies, such as Holzknecht et al. (2018), which have revealed the success of the adaptation of CEFR on a CEFR-aligned language test, Cumming et al. (2002) contend that the assessment of written essays is an interpretive and subjective activity that is heavily influenced by the prevalent educational standards in a particular context. Accordingly, there has been a growing demand for further studies to explore and address the potential impact of flexibility on the test components of CEFR-aligned language tests.

Ongoing amendments have been made to language testing administration and teaching methodology based on the CEFR because CEFR-aligned language tests are developed based on contextual educational standards. Nonetheless, there has been a limited number of recognised organisations that endorse CEFR-aligned language tests to identify the strengths and flaws in each interpretation of the CEFR. Exploring the current variations in CEFR-aligned language tests is important in understanding the potential impacts of the flexibility of CEFR on different aspects of CEFR-aligned language tests. A thorough examination of the methods employed in the tests allows for an in-depth understanding of the strengths, weaknesses, and innovations of the studies conducted on the tests. The findings would have the potential to inform language learners, educators, testing organisations, and policymakers and contribute to the improvement of studies of language testing practices aligned with the CEFR. Accordingly, the current systematic review was driven to answer a research question: What are the current variations in the methodological components of CEFR-aligned language test studies?

LITERATURE REVIEW

The flexibility of the Common European Framework of Reference (CEFR) has enabled testing organisations and examination boards to create and administer language tests that are in accordance with the CEFR. However, these language tests are dependent upon their designated purpose and context. The variability of CEFR-aligned language tests has hindered research on the tests, especially in assessing writing language tests. Such tests pose interpretive and judgmental challenges as they are highly dependent on contextual educational standards. Educational standards specific to each context exert a profound influence on the intricacies of test formats, scoring criteria, and protocols for test administration, which stem from one or a combination of five purposes of language tests, namely aptitude tests, proficiency tests, placement tests, diagnostic tests, and achievement tests (Brown et al., 2004). Due to the inherent characteristics of writing language tests and the variability of CEFR-aligned language tests, there has been a degree of inconsistency in the methodological aspects of CEFR-aligned language research aimed at addressing the strengths and limitations of each test.

To produce findings that accurately reflect the impact and appropriateness of CEFR-aligned language tests, well-thought-out methodological components should be adopted. Given that each research method carries its strengths and weaknesses, careful attention to the methodological components employed in the design process of a study can significantly improve a study's validity and reliability (Abowitz & Toole, 2010). Researchers often begin new studies by duplicating or implementing similar designs to those of previous studies to confirm a study's validity and reliability. However, a singular study cannot be considered conclusive. Therefore, analysing and comparing the findings of multiple studies can aid in making more reliable determinations regarding the efficacy of the methodological components employed within the literature (Bergmann et al., 2018). Considering the varied purpose of language tests studied in CEFR-aligned language tests, it is imperative to recognise the patterns of methodological components in previous studies to produce valid and reliable interpretations (Mendoza & Knoch, 2018). These interpretations may be influenced by the selection of participants and raters, evaluation methods, and data analysis techniques. By meticulously scrutinising these methodological components, the intricate interplay between context, educational policy, and the outcomes of studies on CEFR-aligned language tests can be unravelled.

The expanded understanding of methodological components elevates the discussion regarding CEFR-aligned language tests. This enables researchers to delve into the multifaceted nature and the significant influence of context and educational policy on the methodological components and implementation of studies on CEFR-aligned language tests. By acknowledging these components, researchers can provide invaluable insights into the broader implications and applications of CEFR-aligned language tests across diverse educational settings.

METHODOLOGY

The selected methodology for this study was a systematic review approach, which is appropriate for analysing methodological discrepancies in studies conducted on language tests that are aligned with the Common European Framework of Reference (CEFR). The methodology facilitated the identification of similarities, disparities, and emerging patterns in the studies on CEFR-aligned

language tests, as reported in journal articles, to gain comprehension of present methodological approaches (Polanin et al., 2016).

Search Process
Following the guidelines of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) (Moher, 2009), the search process began with a search on ScienceDirect. The search was conducted between 11 January 2022 and 15 February 2022 for journal articles published between 2011 and 2021. Only one database was utilised in the current review to conduct an efficient search in adherence to Adams et al. (2017), who stated that focusing only on a database allows optimal coverage of the database.

Four main keywords were utilised: CEFR, writing, assessment, and rating scale. The keyword 'assessment' was utilised during the searching process instead of 'test' or 'language test' to vary the findings, considering that assessment is the superset of tests (Brown et al., 2004). Search words included variations of the main keywords, including synonyms and related words, to expand the search and capture a broader range of relevant studies. An online thesaurus, past research keywords, and the Scopus database were consulted to identify additional terms and ensure comprehensive coverage (see Table 1).

Table 1
Identified keywords and their variations

Keyword	Variations
CEFR	Common European Framework of Reference, language proficiency, CEFR-aligned.
Writing	Written Products, writing checklists, writing language test, writing ability, essays, L2 writing, L2 written products, writing skills, writing practices.
Assessment	Assess, scores, validating, examination, test, testing, L2 performance.
Rating scales	Multi-scales, scoring, rating scales, Common European Framework subscale.

Moreover, advanced searches using Boolean logical operators Boolean Operator (AND, OR) and phrase searching were also used to locate related articles in the chosen database (refer to Table 2).

Table 2
Search string for searching database articles.

Search Technique	Search String
Phrase searching	"CEFR assessment" "CEFR writing" "CEFR rating scales"
Boolean Operator	((("CEFR" OR "Common European Framework of Reference") AND ("writing") AND ("assessment") AND ("rating scales"))

In addition, expert opinions were also considered to refine the search strategy and identify potential gaps in the existing literature (Evans & Benefield, 2001).

Eligibility criteria

The selected journal articles met several criteria. First, the journal articles were published between 1 January 2011 and 31 December 2021. The articles had to be published within the last ten years (2011 to 2021) to capture the progress and trends in the CEFR research, which experienced relatively slow progression (Nguyen & Hamid, 2021). Second, the articles should contain relevant empirical data to ensure the inclusion of robust research findings. The selected articles should focus on how CEFR was utilised in the given context. In addition, the journal articles was limited to those written in English to ensure consistency in analysis, given the use of similar terminologies and descriptions in the language. Review articles were also excluded from this review, as the primary focus of the current review was on examining prior research findings rather than evaluating past research assessments. Details of inclusion and exclusion criteria are displayed in Table 3.

Table 3
The inclusion and exclusion criteria.

	Inclusion criteria	Exclusion criteria
Year of publication	Ten years (2011–2021)	<2011
Publication type	Journal articles	Journals (review), book series, book, chapter in a book, conference proceeding
Subject areas	Social Sciences	Other than social sciences
Language	English	Other than English
Types of findings	Empirical	
Focus of findings	Data related to the writing language test / CEFR writing	

Also, the current review employed the population, intervention, comparison, outcomes, and study design (PICOS) to ensure a focused and comprehensive review.

Table 4
The PICOS table

PICOS	Inclusion criteria	Exclusion criteria
Population	Current or aspiring academicians and/or faculty members within educational institutions, research, and/or any academic settings.	Undergraduate and postgraduate educators and researchers
Intervention	Journal articles that addressed the English assessments for current and aspiring academicians and/or faculty members within educational institutions, research, and/or any academic settings. Essential criteria: • Evidence of English assessments and	Documentation methods that does not include personal and/or collective intellectual engagement with the content

	rating scales • Include personal and/or collective intellectual engagement with the content	
Comparison	Articles that addressed the following comparisons were also included: • Comparison of the various CEFR-aligned language tests used in different settings • Comparison of the use of rating scales in different CEFR-aligned language tests • Comparisons between English and CEFR-aligned language tests	
Outcomes	Articles that measured the following outcomes were included: • Impact of the use of rating scales on English assessments • The influence of CEFR on CEFR-aligned language tests • The role of CEFR scales in constructing rating scales	
Study design	Articles that applied: • Qualitative, quantitative, and mixed method Articles in English Year of Publication: 2011–2021 Type of findings: Empirical	

Although the PICOS tools are commonly used in clinical settings, utilising this format in this review could enhance precision and effectively define the parameters of the review (Moher et al., 2009) (see Table 4).

Study Selection
A search query comprising a set of specific keywords executed on the ScienceDirect database retrieved a total of 322 journal articles. Following the inclusion and exclusion criteria process, a total of 284 articles were excluded, leaving only 38 articles eligible for the eligibility process. Of the 38 journal articles, 7 articles were excluded. A total of three journal articles (i.e., Wang et al., 2012; Weigle, 2013; Garner et al., 2019) were deemed suitable for analysis due to their robust research design and clear reference to the marking descriptor, one review paper (i.e., Melissourgou & Frantzi, 2015), one journal article (i.e., Qin & Uccelli, 2020) was excluded due to an unclear number of raters, and two journal articles (i.e., Lee, 2021; Yoon, 2017) were excluded due to a lack of close reference to the marking descriptor. The details of the current study PRISMA flow diagram, which consists of identification, screening, and included processes, are displayed in Figure 1.
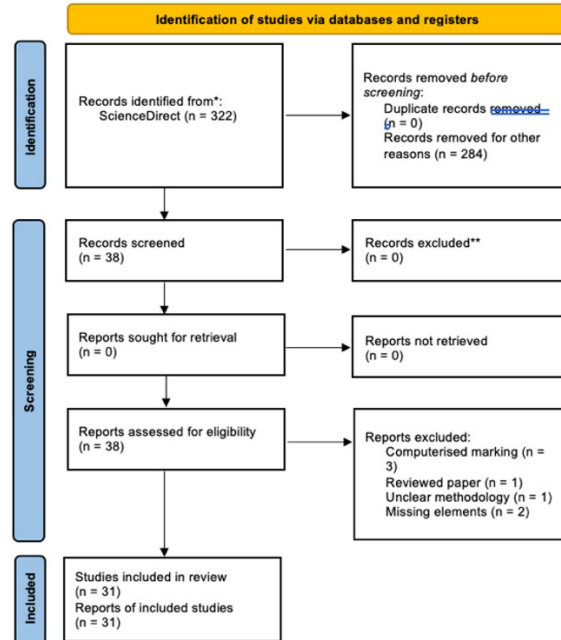
Figure 1: PRISMA 2020 flow diagram (adapted from Page et al., 2021).

Data Extraction and Quality Assessment

A field expert, who was a university lecturer with experience in CEFR studies, was consulted to determine the inclusion of journal articles. When no consensus was reached between the researcher and the field expert, a thorough examination of the methodology, results, and discussion sections was conducted. The journal articles were assessed using the Quality Assessment Rubric, which consisted of seven criteria: Objectives and Purposes, Review of the Literature, Theoretical Frameworks, Participants, Methods, Results and Conclusions, and Significance. The standard of quality reporting employed the Quality Assessment Rubric adapted from Mullet et al. (2017) by Margot and Kettler (2019) (refer to Table 5). In the adapted version, each criterion was evaluated using a four-point scale. Journal articles scoring 14 or less were considered to have failed to meet the quality standard. They were excluded from the final selection. All 31 journal articles assessed received a score higher than 14, indicating their high quality.

RESULTS AND DISCUSSION

The current systematic review gathered 31 journal articles on CEFR-aligned language tests. There were six prominent discoveries: study method, data analysis, participant characteristics, rater participation, number of essays, and writing genres.

Previous journal articles on CEFR-aligned language tests adapted quantitative (QN) and mixed-methods (MX) approaches as their design methods.

Table 5
Study methods

Study Method	Count	Percentage
QN	20	64.5 %
MX	11	35.4 %

The analysis of previous CEFR-aligned language tests reveals a higher prevalence of quantitative approaches (64.5 %) compared to mixed-methods approaches (35.4 %) (see Table 5). The analysis techniques employed in the reviewed journal articles further support the use of this method, with a majority categorised as quantitative analyses, such as MRFM, Pearson Correlation Coefficient, Mann-Whitney U Test, and Likert Scale (see Table 6). According to Table 6, the most frequently utilised analysis in journal articles on CEFR-aligned writing language tests was MRFM, which was mentioned in 6 journal articles, accounting for 19.3 per cent. Other commonly mentioned analyses included Pearson Correlation Coefficient in 3 (9.68 %) journal articles and Exploratory Factor Analysis and Cronbach's alpha in 2 (6.45 %) journal articles. Various other analyses were also adopted, such as G Theory and Likert Scale, each in 2 journal articles (6.45 %).

Table 6
Data analysis in journal articles on CEFR-aligned writing language tests

Analysis	Count	Percentage
G Theory	2	6.45 %
MRFM	6	19.35 %
Pearson Correlation Coefficient	3	9.68 %
Q Matrix	1	3.23 %
Mann-Whitney U Test	1	3.23 %
SPSS Program & AA-ICC Two-way Mixed Method	1	3.23 %
LCA, TAALED, Coh-Metrix	1	3.23 %
Exploratory Factor Analysis, Cronbach's alpha	2	6.45 %
Likert Scale	2	6.45 %
Text Inspector & Human Analyst	1	3.23 %
Group interview, questionnaire, Rasch analysis	1	3.23 %
CHAT, CLAN, M/ANOVA	1	3.23 %
Coh-Metrix & AVA	1	3.23 %
ANOVA discussion	1	3.23 %
Facets, R	1	3.23 %
WCR, L2SCA	1	3.23 %
FACETS	1	3.23 %
Questionnaire	1	3.23 %
R scripts	1	3.23 %
Cut scores, Standard setting	1	3.23 %

CVI	1	3.23 %
Syntactic Elaboration Model, ANOVA	1	3.23 %
C-Test, Pearson Correlation	1	3.23 %
METOOL, LCEUPV	1	3.23 %
Interview	1	3.23 %
Edu-G	1	3.23 %
No specific analysis mentioned	1	3.23 %

Note: One study (Study [19]) does not mention any specific analysis.

This indicates a preference for quantitative research in the field and is against scholars such as Xi and Sawaki (2017). Xi and Sawaki (2017) believe that quantitative methods are valuable for testing hypotheses, hence they may fall short in generating new hypotheses and providing rich insights into the intricate nature of language testing. The high percentage of the use of quantitative methods in CEFR-aligned studies could have been due to the nature of most studies focusing on rubric development, particularly in examining the correlation between test-taker proficiencies and rubric descriptors, often utilising Rasch measurement (Janssen et al., 2015).

The absence of a sole qualitative design in the provided information could be attributed to the measurement requirements in the field of language testing (McNamara & Knoch, 2012). Nevertheless, this does not diminish the value or use of qualitative research in the field. Some reviewed articles employed qualitative analyses, including interviews, Q Matrix, and G-Theory, often assumed to be part of mixed-methods approaches. Nonetheless, Fox (2017) recognises the widespread use of qualitative and mixed methods in language testing studies. Qualitative research supports the descriptive the nature of language testing, while quantitative methods are associated with tests and psychometrics. However, the prevalence of mixed methods approaches is growing. More articles have combined quantitative and qualitative methods to address issues in language testing. This integration allows for a more comprehensive understanding of language assessment, leveraging the strengths of both approaches. The increasing use of mixed methods reflects the recognition of the value of integrating different perspectives and approaches that could lead to more robust findings in language testing research. In general, there has been a higher prevalence of quantitative methods in the studies of CEFR-aligned language tests while recognising the importance of qualitative and quantitative approaches.

The studies on CEFR-aligned language tests were conducted on participants with different characteristics. Previous reviews have revealed an expansive range of sample sizes and age groups among participants in language learning engagement research. The current review discovered that a significant 67.74 per cent of the analysed journal articles chose participants based on their language proficiency level (refer to Table 7).

Table 7
The participants' language proficiency levels

Language Level	Count	Percentage
A1 – B1	3	10 %
A1 – C2	1	3.33 %

A2 – B1	3	10 %
A2 – B2	3	10 %
B1 – B2	3	10 %
B1 – C1	1	3.33 %
B2 – C2	4	13.33 %
C2	1	3.33 %

This focus on proficiency indicates a clear tendency in language learning engagement research towards studying participants with certain language capabilities, specifically those in the B2–C2 range (13.33 %). This observation supports Melissourgou and Frantzi's (2015) assertion that B2-level certifications are highly sought after in professional settings, with C2 necessary for roles demanding advanced language expertise. Another reason could be that general English writing papers are designed at the B2 level (Lukácsi, 2021). For example, in Belgium, two of the country's CEFR-aligned language tests; namely, STRT (Educatief Startbekwaamand) and ITNA (Interuniversitaire Taaltest Nederlands voon Anderstaligen), have been formally linked to the B2 level of the CEFR (Deygers et al., 2018).

The focus on participants within the B2–C2 proficiency range in CEFR-aligned language studies reflects a deliberate consideration of the criteria for participant selection. These studies prioritise individuals who possessed a higher level of language proficiency, as demonstrated by their placement within the B2–C2 range. The choice to study participants aligns with the specific requirements of professional settings, where B2-level certifications are highly sought after (Melissourgou & Frantzi, 2015). Furthermore, the inclusion of B2-level design in general English writing papers (Lukácsi, 2021). For example, in Belgium, two CEFR-aligned language tests, namely STRT (Educatief Startbekwaamand) and ITNA (Interuniversitaire Taaltest Nederlands voon Anderstaligen) have been officially linked to the B2 level of the CEFR (Deygers et al., 2018). This pattern reinforces the tendency of selecting participants within the B2–C2 proficiency range. In short, the focus on participants with B2–C2 proficiency in CEFR-aligned language studies reflects a deliberate consideration of the criteria for participant selection. By examining individuals within this proficiency range, researchers aim to represent the language abilities required in professional settings accurately, indicated by B2-level certifications and general English writing papers. In addition, the selection of participants of CEFR-based language tests was based on the academic level of participants. The current review found 29 per cent of the article used this criterion as the basis for their selection. Among these, undergraduate students were most frequently represented, making up 13.33 per cent of the total (refer to Table 8).

Table 8
The participants' academic level

Academic Level	Count	Percentage
Undergraduate	4	13.33 %
Master's Degree	1	3.33 %
Graduate	1	3.33 %
University students	1	3.33 %
3rd grade	1	3.33 %
Secondary school students	1	3.33 %

The emphasis on selecting undergraduate students as research participants aligns with the findings of Mendoza and Knoch (2018), who highlight their superior writing skills, making them well-suited for writing tests. This preference can also be attributed to the fact that numerous CEFR-aligned language tests have been specifically developed for undergraduate students. These tests play a crucial role in assessing their language proficiency and ensuring they meet the necessary standards. In China, the English proficiency test (EPT) is a mandatory exam for undergraduate students to complete their studies. Meanwhile, the College English Test Band 4 (CET-4) is China's nationally recognised assessment that evaluates the English language skills of non-English major undergraduate students. Its purpose is to ensure that these students meet the established standards of English proficiency (Winke & Lim, 2017). Thus, the alignment between the preference for undergraduate participants and the abundance of tailored language tests further underscores the significance of this demographic in studies pertaining to CEFR-aligned language tests.

The variability of raters in CEFR-aligned writing language tests was significant. As a result, a central point of discussion revolved around determining the optimal number of raters required to ensure the reliability of test results. The reviewed journal articles displayed a range of rater involvement, with the number of raters varying from 2 to 210. Due to the substantial variation, the data were categorised into three distinct groups: low, medium, and high (refer to Table 9). Upon examination of Table 9, the majority of the journal articles were classified as either low or high in terms of the number of raters. Specifically, 12 journal articles, which account for 38.71 percent, utilised a limited number of raters, ranging from 2 to 7, and the other 12 journal articles, which also account for 38.71 per cent, employed a more substantial number of raters, ranging from 35 to 210.

Table 9
Categories for the number of raters and number of essays

Category	Number of Raters	Count	Percentage	Number of Essays	Count	Percentage
Low	2 – 7	12	38.71 %	50 – 99	2	6.45 %
Medium	8 – 17	7	22.58 %	100 – 360	9	29.03 %
High	35 – 210	12	38.71 %	410 – 5236	20	64.52 %
Total		31	100 %		31	100 %

Likewise, the range of the number of essays evaluated in CEFR-aligned writing language tests was also found to be significant. According to Table 9, the number of essays varied from 50 to 5236. To account for this substantial variation, the number of essays was also categorised into three distinct groups: low, medium, and high (refer to Table 9). Despite claims that handling such a high volume of essays within a limited timeframe presents additional challenges related to training and maintaining marker confidence and reliability (Brown et al., 2004), the current review discovered that 64.52 per cent of journal articles fell into the high category, assessing a significant number of essays ranging from 410 to 5236. This indicates a clear tendency towards conducting large-scale investigations. According to Brown et al. (2004), the deliberate efforts to gather extensive data and capture a wide range of perspectives is to strengthen the robustness and reliability of study findings.

These large-scale studies serve multiple purposes in the field of CEFR-aligned language tests. Firstly, they aim to simulate high-stakes testing conditions, replicating the contexts where

the consequences of test performance can have significant implications for test takers (Fleckenstein et al., 2020). By conducting large-scale studies, researchers can better understand and evaluate the impact of CEFR-aligned language tests in such high-stakes contexts. Secondly, these studies account for the presence of relatively large individual differences among test takers, such as variations in language proficiency and background in high-stakes contexts (Deygers et al., 2017). Including a diverse range of participants in large-scale studies allows researchers to capture and account for these individual differences, leading to a more comprehensive understanding of the effectiveness and applicability of CEFR-aligned language tests. In short, the decision to employ a higher number of raters and assess a larger quantity of essays in these studies reflects the conscientious efforts made to ensure the robustness and generalisability of the findings (Harsch & Rupp, 2011). For example, the large-scale studies conducted by Brown et al. (2004), Deygers et al. (2017), and Harsch and Rupp (2011) have been essential for generalising the appropriateness of CEFR-aligned language tests in specific contexts. Thus, the tendency towards large-scale studies in CEFR-aligned language tests is evidenced by the inclusion of a high number of raters and essays. These studies simulate high-stakes testing conditions, account for individual differences among test takers, and contribute to the validity, reliability, and generalisability of the findings.

The significance of different writing genres in CEFR-aligned language test studies cannot be overlooked because the choice of genre impacts the complexity and quality of written output, ultimately influencing the determination of CEFR levels. Previous research has demonstrated the impact of genres on L2 writing production, showcasing variations in linguistic features, cognitive demands, and communicative functions across different genres. Thus, the current review includes insights into the variety of writing genres employed in CEFR-aligned writing language tests. A total of 83.87 per cent of the reviewed journal articles adopted a specific writing genre.

Table 10
Essay types in journal articles on CEFR-aligned writing language tests

Genre	Frequency	Percentage
Reading-to-Write (RTW)	2	6.45 %
Opinion essay	7	22.58 %
Letter/Composition	1	3.23 %
Descriptive	2	6.45 %
Dissertation	1	3.23 %
Response Essay	1	3.23 %
Academic writing	1	3.23 %
Email/Blog	2	6.45 %
Narrative	2	6.45 %
General Essay	1	3.23 %
Argumentative Essay	5	16.13 %
Integrated & Independent Writing	1	3.23 %
All four skills	4	12.90 %
Not mentioned	1	3.23 %
Total	31	100.00 %

Table 10 presents a breakdown of the essay types in journal articles on CEFR-aligned writing language tests along with their corresponding frequencies and percentages. The variety of the genre adopted can be influenced by the proficiency and academic level of the participants. According to Neff-van Aertselaer (2013), the choice to utilise a genre in language tests is rooted in the recognition that students' writing skills progress from explanatory to argumentative genres. Lee (2021) concurs that genres such as descriptive essays or basic narratives are linked to lower CEFR levels because they require less complex language use and have lower cognitive demands.

According to Table 10, the most common writing genre in the journal articles on CEFR-aligned writing language tests was the opinion essay, with a frequency of 7 (22.58 %) articles. The high tendency to adopt the opinion writing genre when studying CEFR-aligned tests attends to the general level of proficiency of participants in the study, which was participants with B2–C2 proficiency. According to Harsch and Rupp (2011), tasks at higher CEFR levels, such as B2, require more elaborate responses, such as writing an opinion on a topic of general interest. Despite the higher tendency for studies on CEFR-aligned language tests to employ the opinion genre, there have been solid arguments to have used the argumentative genre to explain the 16.15 per cent use of the argumentative genre in studies on CEFR-aligned language tests. First, the genre requires moderate to advanced language skills and sophisticated reasoning abilities, thus attending to the level of most study participants, B2–C2 proficiency. Barrot and Agdeppa (2021) denote argumentative essays as an index of proficiency to study the interaction between language proficiency and CAF measures of the International Corpus Network of Asian Learners of English (ICNALE). Second, the argumentative genre has been frequently adopted in CEFR-aligned language tests, including Selectividad, a CEFR-aligned Spanish university entrance examination (Neff-van Aertselaer, 2013) and TOEFL iBT (Fleckenstein et al., 2020). In general, prior articles written on CEFR-aligned language tests considered the genre selection, with a predominant use of the opinion writing genre. This aligns with the proficiency levels of participants in the B2–C2 range and the expectations of higher CEFR levels. However, the presence of the argumentative genre highlights its relevance for participants with moderate to advanced language skills and sophisticated reasoning abilities.

CONCLUSION

The current systematic review employed a comprehensive search strategy using specific keywords and inclusion/exclusion criteria to identify relevant articles published between 2011 and 2021 on CEFR-aligned language tests. The search process yielded 322 journal articles, and after screening and assessing their quality, 31 articles were included in the review. The current review thoroughly analysed the selected journal articles on CEFR-aligned language tests to provide a framework for replicating previous studies and serve as a reference for future studies.

The review identified six tendencies of methodological patterns in the studies of CEFR-aligned language tests: research methods, participant characteristics, sample sizes, rater numbers, essay quantities, and writing genres. In short, the findings revealed a preference for quantitative research methods in the field, although there was recognition of the value of qualitative and mixed-methods approaches. Prior studies have also mostly focused on participants within the B2–C2 proficiency range, and there is a higher tendency for studies to be conducted in professional

settings examining language tests for undergraduate students. In addition, efforts were made to ensure the robustness and generalisability of the findings by including larger sample sizes and more raters and essays, particularly in high-stakes testing contexts. The choice of writing genres in CEFR-aligned language tests was found to impact the complexity and quality of written output and the determination of CEFR levels. Overall, to enhance the replicability of studies on CEFR-aligned language tests in order to increase study validity and reliability, it is highly recommended to employ a combination of quantitative and mixed-methods approaches, utilise appropriate data analysis techniques, recruit participants within the B2–C2 proficiency range at the undergraduate academic level, involves a large number of raters and written essays, and integrate the opinion writing genre to reflect the language proficiency of test takers.

REFERENCES

Abowitz, D. A., & Toole, T. M. (2010). Mixed method research: Fundamental issues of design, validity, and reliability in construction research. *Journal of Construction Engineering and Management, 136*(1), 108–116. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000026

Adams, R. J., Smart, P., & Huff, A. S. (2017). Shades of grey: Guidelines for working with the grey literature in systematic reviews for management and organizational studies. *International Journal of Management Reviews, 19*(4), 432–454.

Barrot, J. S., & Agdeppa, J. Y. (2021). Complexity, accuracy, and fluency as indices of college-level L2 writers' proficiency. *Assessing Writing, 47*, 100510. https://doi.org/10.1016/j.asw.2020.100510

Bergmann, C., Tsuji, S., Piccinini, P. E., Lewis, M. L., Braginsky, M., Frank, M. C., & Cristia, A. (2018). Promoting replicability in developmental research through meta-analyses: Insights from language acquisition research. *Child Development, 89*(6), 1996–2009.

Brown, G. T., Glasswell, K., & Harland, D. (2004). Accuracy in the scoring of writing: Studies of reliability and validity using a New Zealand writing assessment system. *Assessing Writing, 9*(2), 105–121. https://doi.org/10.1016/j.asw.2004.07.001

Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *The Modern Language Journal, 86*(1), 67–96.

Deygers, B., Van den Branden, K., & Peters, E. (2017). Checking assumed proficiency: Comparing L1 and L2 performance on a university entrance test. *Assessing Writing, 32*, 43–56. https://doi.org/10.1016/j.asw.2016.12.005

Deygers, B., Van den Branden, K., & Van Gorp, K. (2018). University entrance language tests: A matter of justice. *Language Testing, 35*(4), 449–476.

Evans, J., & Benefield, P. (2001). Systematic reviews of educational research: Does the medical model fit? *British Educational Research Journal, 27*(5), 527–541.

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

Fox, J. (2017). Using portfolios for assessment/alternative assessment. In E. Shohamy, I. G. Or, S. May, & E. Shohamy (Eds.), *Language testing and assessment* (pp. 135-148). Springer.

Harsch, C., & Rupp, A. A. (2011). Designing and scaling level-specific writing tasks in alignment with the CEFR: A test-centered approach. *Language Assessment Quarterly, 8*(1), 1–33.

Holzknecht, F., Eberharter, K., Größinger, J., & Kremmel, B. (2018). Using the CEFR Companion Volume for mapping workplace English needs. *Language Value, 10*(1), 28–54.

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

Lee, J. (2021). Using corpus analysis to extend experimental research: Genre effects in L2 writing. *System, 100,* 102563. https://doi.org/10.1016/j.system.2021.102563

Lukácsi, Z. (2021). Developing a level-specific checklist for assessing EFL writing. *Language Testing, 38*(1), 86–105. https://doi.org/10.1177/0265532220916703

Margot, K. C., & Kettler, T. (2019). Teachers' perception of STEM integration and education: A systematic literature review. *International Journal of STEM Education, 6*(1), 1-16. https://stemeducationjournal.springeropen.com/articles/10.1186/s40594-018-0151-2

Mat Yusoff, S., Arepin, M., & Mohd Marzaini, A. F. (2022). Secondary school teachers' perspectives towards the implementation of CEFR-Aligned English Curriculum. *Creative Practices in Language Learning and Teaching, 10*(1), 32–48.

McNamara, T., & Knoch, U. (2012). The Rasch wars: The emergence of Rasch measurement in language testing. *Language Testing, 29*(4), 555-576. https://doi.org/10.1177/0265532211430367

Melissourgou, M. N., & Frantzi, K. T. (2015). Testing writing in EFL exams: The learners' viewpoint as valuable feedback for improvement. *Procedia - Social and Behavioral Sciences*, *199*, 30–37. https://doi.org/10.1016/j.sbspro.2015.07.483

Mendoza, A., & Knoch, U. (2018). Examining the validity of an analytic rating scale for a Spanish test for academic purposes using the argument-based approach to validation. *Assessing Writing, 35,* 41–55. https://doi.org/10.1016/j.asw.2017.12.003

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., The PRISMA Group. (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLoS Med 6*(7): e1000097. https://doi.org/10.1371/journal.pmed.1000097

Mullet, D. R., Rinn, A. N., & Kettler, T. (2017). Catalysts of women's talent development in STEM: A systematic review. *Journal of Advanced Academics, 28*(4), 253-289.

Nagai, N. (2020). A critical look at the validation of Cambridge English Writing Exams: Experts' evaluation of the CEFR theory of action. *Language Testing, 37*(2), 292–311.

Neff-van Aertselaer, J. (2013). Contextualizing EFL argumentation writing practices within the Common European Framework descriptors. *Journal of Second Language Writing, 22*(2), 198–209. https://doi.org/10.1016/j.jslw.2013.03.010

Nguyen, V. H., & Hamid, M. O. (2021). The CEFR as a national language policy in Vietnam: Insights from a sociogenetic analysis. *Journal of Multilingual and Multicultural Development, 42*(7), 650–662. https://doi.org/10.1080/01434632.2020.1715416

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ... & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*(71). https://doi.org/10.1136/bmj.n71

Polanin, J. R., Maynard, B. R., & Dell, N. A. (2016). Overviews in education research: A systematic review and analysis. *Review of Educational Research, 87*(1), 172–203.

Winke, P., & Lim, H. (2017). The effects of test preparation on second-language listening test performance. *Language Assessment Quarterly, 14*(4), 380–397.

Xi, X., & Sawaki, Y. (2017). Methods of test validation. In E. Shohamy, I. G. Or, & S. May (Eds.), *Language testing and assessment* (pp. 193–210). Springer.

Acknowldgement

Conflict of Interest

We declare no conflicts of interest regarding the research titled 'Variations and Methodological Components in CEFR-Aligned Language Tests: A Systematic Review'. Our involvement in this research is solely for academic purposes, and we do not have any financial or personal relationships that could potentially bias or influence the outcomes of this study.

Authors' Contributions

Nik Aloesnita Nik Mohd Alwi significantly contributed to the conceptualisation, methodology design, data curation, and drafting of the original manuscript. Wan Alisa Hanis Wan Abdul Halim conducted the literature review, performed investigation tasks, conducted formal analysis, and contributed to the manuscript's review and editing process. Both authors have contributed equally to this research, ensuring its comprehensive development and completion.

About the Authors

	Nik Aloesnita Nik Mohd Alwi, *PhD*, has over 20 years of experience at Universiti Malaysia Pahang Al-Sultan Abdullah. Her research and publications are mainly in second language acquisition, task-based learning, and technology-enhanced language learning including CEFR-related. She contributes her academic and research expertises through seminars, workshops, and involvement with the Malaysian Qualifications Agency.
	Wan Alisa Hanis Wan Abdul Halim, a language teacher currently pursuing an MSc at Universiti Malaysia Pahang Al-Sultan Abdullah, has a decade of expertise in the field. Possessing specialised expertise in national-level English language assessment, she is actively contributing to the execution, and analysis of local English language assessments. Her involvement extends to impactful CEFR-related research initiatives at the academic level.