

---

## A Systematic Review on Argumentative Writing Evaluation Frameworks in Cefr-Aligned Esl Contexts

Vicknes Gopal Krishnan<sup>\*1</sup>, Afendi Hamat<sup>2</sup>, Anis Nadiah Che Abdul Rahman<sup>3</sup>

<sup>1,2,3</sup>Centre for Research in Language and Linguistics (PKBL), Faculty of Social Sciences and Humanities,  
Universiti Kebangsaan Malaysia

<sup>1</sup>[P151782@siswa.ukm.edu.my](mailto:P151782@siswa.ukm.edu.my), <sup>2</sup>[fendi@ukm.edu.my](mailto:fendi@ukm.edu.my) <sup>3</sup>[aniscar@ukm.edu.my](mailto:aniscar@ukm.edu.my)

**\*Corresponding Author**

Vicknes Gopal Krishnan

Centre for Research in Language and Linguistics (PKBL), Faculty of Social Sciences and Humanities,  
Universiti Kebangsaan Malaysia

Email ID: [P151782@siswa.ukm.edu.my](mailto:P151782@siswa.ukm.edu.my)

### ABSTRACT

The study of argumentative writing and argumentative strength in ESL settings has grown significantly over the last two decades. However, the literature is still disjointed with respect to theoretical orientation, operational definitions and methodology. Specifically, the systematic synthesis of the analytic argument models, rubric-based assessment frameworks, and computational models applied to measure the strength of arguments in ESL writing among students are lacking. This systematic literature review integrates existing studies on the evaluation of argumentative strength in ESL writing, focusing on how argumentation is conceptualized, operationalized and assessed across contexts. Following PRISMA guidelines, qualitative research synthesis was used to analyse peer-reviewed studies published between 2010-2024. Findings highlights trends in evaluation frameworks, operational definitions of argumentative strength, methodological approaches and existing limitations. Analytical-model frameworks, particularly those based on Toulmin's model, provide more explicit and transparent representations of argumentative reasoning for diagnostic and explanatory purposes. In contrast, rubric-based and computational approaches offer less detail regarding internal reasoning processes. The review also identifies limited alignment between CEFR writing descriptors and structural argumentation models, as well as a shortage of research conducted in the Malaysian ESL context. Implications for future research and assessment practices are discussed.

**Keywords:** systematic review; argumentative writing; ESL writing assessment; Toulmin's model; argumentative strength.

### INTRODUCTION

It is commonly acknowledged that argumentative writing has become one of the fundamental aspects of academic literacy since argumentative writing requires a learner to come up with the claims, validate the point using sound arguments, and address opposing perspectives in a logical and compelling way. In the ESL environment, the capability of constructing a successful argument has been linked to the superior skills in higher-order knowledge, academic readiness, and academic achievement in high-stakes examinations, especially at the secondary and tertiary levels (Heng et al., 2023; Morris, Chan, and Khoo, 2024). As a result, the topic of argumentative writing is now a point of interest in the study of second language writing, and extended attention is paid to the conceptualisation and evaluation of argumentative strength. Regardless of this accumulating literature, studies that investigate argument strength with respect to ESL writing indicate a big discrepancy in the assessment systems used. The current literature is based on various approaches that vary in terms of theoretical basis, depth of assessment and rigour of methods used. Other studies have applied analytic models that are based on argumentation theory in order to analyse the internal organisation of reasoning, whereas others have utilised rubric based writing tests in harmony with standardised tests or proficiency models. Numbers Computational and automated methods have been also more recently introduced to find argumentative features based on linguistic or statistical features (Shen and Yao, 2022; Balkir, Ozturk, and Aydin, 2023). This variety of methods has led to inconsistent definitions of argumentative strength, and different coding and scoring practices, and impedes comparability across studies leading to difficulties to make cumulative findings on argumentative competence of ESL learners. The extensive adoption of the Common European Framework of Reference (CEFR) into ESL teaching has only made the assessment of argumentative strength less explicit and conceptually underspecified. Despite the use of CEFR to offer general descriptions of writing proficiency at any level, the aspect of

argumentation is treated implicitly, and has little to inform the structural elements of effective reasoning such as claims, justification, and counterargumentation (Ahmad and Kabilan, 2021; Hajar and Mahadi, 2022). Consequently, CEFR-based writing rubrics tend to focus more on overall communicative effectiveness, fluency, and coherence and give little analytical assistance in assessing the quality of argumentative reasoning. This constraint is especially acute in ESL examination-based environments, such as the Malaysian education system, in which the learners are expected to write opinion-based pieces which implicitly require a structured form of argumentation, without specifying assessment criteria of the quality of reasoning (Uri, 2021; Thiagarajan and Hua, 2023). With such conceptual and methodological difficulties, a systematic overview of the current literature on studying argumentation strength in ESL writing is urgently needed to map the variability of argumentation frameworks employed in different methods, to compare the operationalisation of argumentation across methodologies, and to define some methodological and contextual gaps. In line with this, the systematic literature review will be directed by the following research question: What is the current situation in the research on the frameworks of argumentative strength evaluation in the context of ESL student writing?

This review contributes to the field by providing a structured and systematic synthesis of the diverse evaluative frameworks used to assess argumentative strength in ESL writing. By critically examining analytical-model, rubric-based and computational-approaches, the study seeks to clarify conceptual inconsistencies and highlight the implications of divergent assessment practices. The significance of this review extends to multiple stakeholders. It offers clearer insights into how argumentative reasoning can be assessed explicitly and effectively beyond general language proficiency for educators. For researchers on the other hand, it highlights the need for greater conceptual alignment and methodological consistency across studies which will then support the development of more robust and comparable research designs. Significantly, within the Malaysian ESL context, where CEFR-aligned assessment has been implemented, this review functions as timely evidence to inform policy, curriculum design and the development of more valid and context-sensitive writing assessment framework.

## **METHODS**

### **The Review Protocol: PRISMA**

To maintain transparency, methodological rigour and reproducibility of the review process, this study relied on the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines. PRISMA gives a systematic approach to the identification, screening, eligibility, and synthesis of the relevant literature in a systematic review. PRISMA has been extensively applied in the research of applied linguistics and writing in the second language to improve the area of consistency in the process of review and reporting. In this regard, the current review adhered to the PRISMA principles when reporting the search strategy, study selection procedures, and data extraction procedures. PRISMA was selected as the guiding framework for this review as the study aims to systematically map, compare and synthesise diverse evaluative framework used to assess argumentative strength in ESL writing rather than to estimate effect sizes. Given the conceptual heterogeneity of argumentation research, PRISMA offers a transparent and a replicable structure for documenting search decisions, inclusion criteria and analytical scope. It highlights that explicit reporting is particularly appropriate for reviews that synthesis methodological and theoretical approaches as commonly undertaken in applied linguistics and writing assessment research (Booth et al., 2021; Xiao et al., 2021)

### **Sources**

The primary literature search was conducted using Scopus and Web of Science, selected for their rigorous indexing standards and comprehensive coverage of peer-reviewed research in applied linguistics, writing assessments and second language studies. Google Scholar was used as a supplementary search tool to support citation chasing and to identify potentially relevant studies not captured by indexed databases. Searches conducted via Google Scholar were restricted to peer-reviewed journal articles with theses, conference proceedings, reports and non-academic sources excluded. Final inclusion decisions were governed by the same eligibility criteria applied to Scopus and Web of Science records to ensure methodological consistency and replicability (Gausenbauer & Haddaway, 2020).

### **Study Selection**

The study selection was carried out in three systemic phases namely; identification, screening, and eligibility evaluation. The search commenced with the identification of possible studies by using database searches and by searching manually. All the retrieved records were then filtered under specified inclusion and exclusion criteria.

The last step was to evaluate the entire text eligibility of the rest of the studies to help in deciding whether they were relevant to the topic of the review.

### Identification

The identification process was initiated by the development of the major search terms associated with the topic of argumentative writing assessment and evaluation systems in the ESL settings. These were combinations of argumentative writing, argument strength, argumentation model, Toulmin, ESL writing assessment, CEFR, writing rubrics and automated or computational assessment. Variations in terminology variations between studies were captured by using Boolean operators (AND, OR) and truncation. The search of the electronic database gave an overall 623 records. Subsequent manual search of reference lists and key journals led to an extra 41 records making the total records identified to 664. 2010-2024 as the time frame was selected to capture progress and developments following the increased international adoption of CEFR informed curricula and assessment framework while acknowledging the growing scholarly attention to argumentation, critical thinking and construct validity in second language writing assessment. To justify further, research prior to 2010 was excluded as it largely predates these conceptual and policy shifts and often treats argumentation implicitly within general writing proficiency measures.

### Screening

During the screening stage, all identified studies were examined based on title and abstract review. Duplicate records were removed, and studies were screened using the following exclusion criteria:

- (a) publications outside the 2010–2024 timeframe,
  - (b) studies not focused on ESL or EFL writing contexts,
  - (c) studies that did not address argumentative writing or argument evaluation,
  - (d) non-peer-reviewed publications such as editorials, book reviews, or opinion pieces, and
  - (e) studies focusing solely on language accuracy without reference to argumentation or reasoning.
- After applying these criteria, 487 studies were excluded, leaving 177 studies for further assessment.

### Eligibility

In the eligibility stage, the remaining 177 studies were subjected to full-text review to ensure alignment with the objectives of the systematic review. Studies were excluded at this stage if they:

- (a) did not explicitly examine frameworks or models for evaluating argumentative strength,
- (b) focused on first-language (L1) writing without ESL relevance,
- (c) addressed general writing assessment without reference to argumentation, or
- (d) lacked sufficient methodological detail regarding evaluation procedures.

Following full-text assessment, 123 studies were excluded, resulting in a final corpus of 54 studies deemed eligible for inclusion in the review.

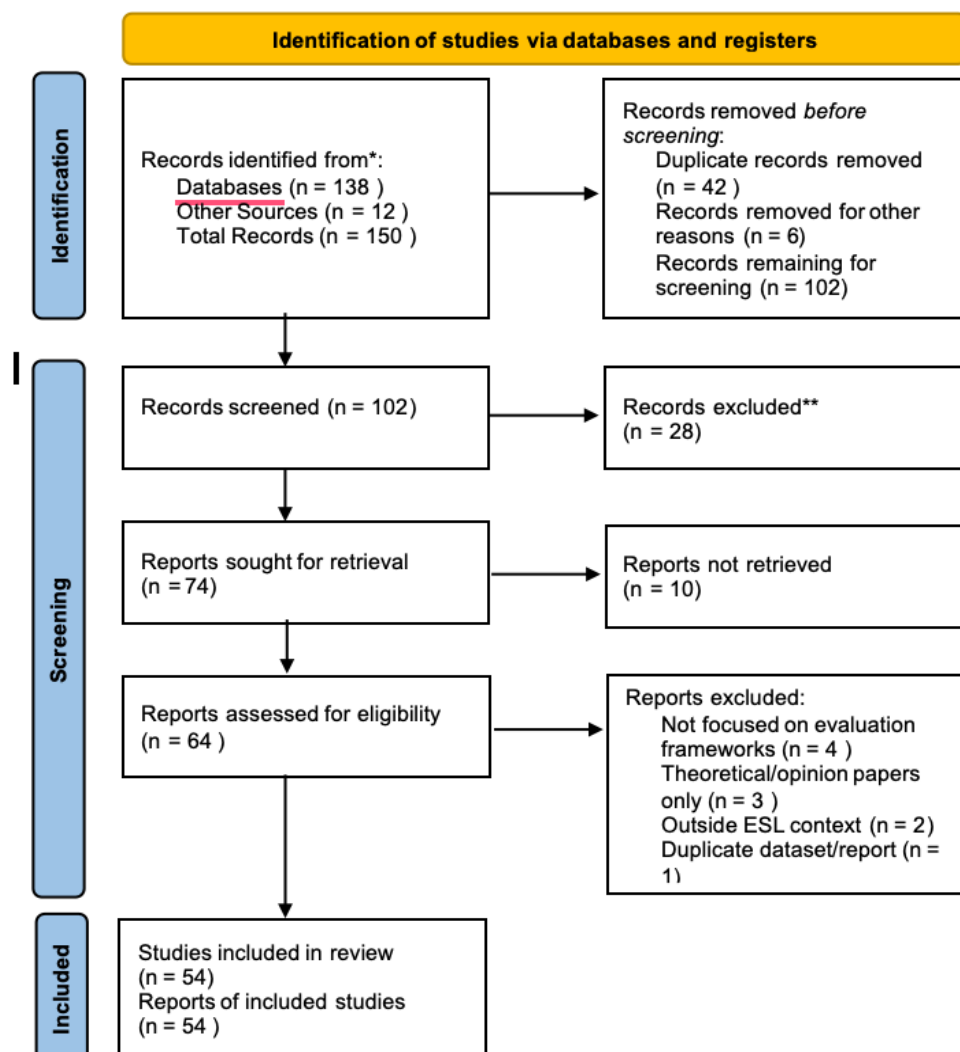
### Data Extraction and Analysis

The final 54 studies were retained for systematic data extraction and qualitative synthesis. These studies constitute an analytically meaningful dataset as the review prioritized conceptual relevance and methodological transparency over numerical extent. Accommodating the focused scope on argumentation strength evaluative framework in ESL writing, the included studies provided sufficient theoretical saturation to enable systematic comparisons across framework types, operational definitions and methodological practices, consistent with established guidance for qualitative research synthesis (Donato & McCormick, 2021). Data were extracted using a structured coding framework that captured the following characteristics:

- (a) research context and learner level,
- (b) theoretical framework underpinning argument evaluation,
- (c) type of argumentative evaluation framework employed (analytic-model, rubric-based, or computational),
- (d) operationalisation of argumentative strength,
- (e) methodological design and data analysis procedures, and
- (f) reported limitations and contextual considerations.

The extracted data were analysed thematically following a qualitative synthesis approach. Initial coding focused on identifying recurring patterns and conceptual categories related to argumentative evaluation practices. These categories were then refined through iterative comparison across studies to generate higher-order themes. The

analysis procedures followed established qualitative synthesis principles and were consistent with thematic analysis approaches commonly applied in systematic reviews within applied linguistics. Through this process, five overarching thematic dimensions emerged, forming the basis for the presentation of findings



As illustrated in Figure 1, the initial search generated 664 records reflecting the broad and interdisciplinary nature of argumentation and writing assessment research. The significant reduction during the screening and eligibility stages highlights the strict application of inclusion criteria particularly the requirement for explicit engagement with argumentative evaluation frameworks. This consecutive filtering ensured that only studies with direct methodological relevance to the review questioned were retained thereby strengthening the internal coherence and defensibility of the synthesis.

### Findings

This section presents the findings of the systematic literature review based on qualitative synthesis of the selected studies. The findings are organised thematically to reflect trends in research focus, types of argumentative strength evaluation frameworks, approaches to operationalising argumentative strength, methodological practices, and identified limitations and contextual gaps in ESL argumentative writing assessment research.

### Trends in ESL Argumentative Writing Assessment Research

**Table 1 Descriptive Overview of Studies on Argumentative Strength Evaluation in ESL Writing (n=54)**

| Category / Framework     | Description                                                       |
|--------------------------|-------------------------------------------------------------------|
| Publication period       | 2010–2024                                                         |
| Research contexts        | ESL (Asia, Europe, Middle East); limited Malaysian studies        |
| Learner levels           | Secondary, pre-university, undergraduate                          |
| Writing genres           | Argumentative essays, opinion essays, persuasive academic writing |
| Dominant research design | Qualitative and mixed-methods                                     |
| Data sources             | Student essays, exam scripts, classroom writing tasks             |
| Main evaluation focus    | Argument structure, reasoning quality, justification              |
| Reliability reporting    | Common in analytic frameworks; limited in rubric-based studies    |

The review of the literature published within the past decade (2010-2024) has shown that there is a definite tendency towards an increase in the number of studies that provide the analysis of the topic of argumentative writing in the ESL learning environment since 2016. Previous literature was inclined to assess the quality of the argumentation indirectly by using holistic writing scales or general proficiency measures; frequently, argumentation was incorporated in bigger constructs, such as coherence or task performance (Lee and Deakin, 2016). Nevertheless, the recent studies reveal a clear change in the focus towards investigating the inner fabric of the argument, which is an indication of a greater awareness of argumentation as a particular and quantifiable part of academic writing proficiency (Wingate, 2021; Liu and Stapleton, 2023). This change has been linked to the increasing concerns on critical thinking, reasoning, and disciplinary literacy in second language learning. Although this is the case, the review states that the growth in research has been without coalescence on a common evaluative framework. Rather, the literature uses a variety of theoretical perspectives and evaluation tools, and thus, the definition, units of analysis, and reporting practices are highly heterogeneous (Shen and Yao, 2022; Morris et al., 2024). Consequently, the results of the research findings are hard to compare in studies, and the cumulative knowledge about the argumentative competence of ESL learners is still discontinuous. Such non-standardisation is especially clear in situations based on CEFR where the argumentative requirements are not explicitly defined but are represented implicitly as a reflection of the overall structure (Hajar and Mahadi, 2022), supporting the importance of systematic comparison of existing systems.

Taken together, the descriptive patterns summarised in Tables 1 to 5 reveal a systematic imbalance between the complexity of argumentative reasoning expected in ESL academic writing and the evaluation mechanisms commonly employed to assess it. While the reviewed studies demonstrate increasing scholarly attention to argumentative writing across diverse ESL contexts (Table 1), the predominance of rubric-based and efficiency-driven assessment practices (Table 2) suggests that argumentative strength is frequently evaluated through indirect or underspecified criteria. This is further evidenced by inconsistencies in how argumentative strength is operationalised across framework types (Table 3), with analytic-model approaches offering explicit structural definitions, while rubric-based and computational frameworks rely on holistic or proxy indicators. Methodological comparisons (Table 4) indicate that although analytic approaches are more labour-intensive, they provide greater transparency and diagnostic value, which is essential for understanding how reasoning is constructed rather than merely judged. Moreover, the research gaps identified in Table 5 highlight persistent misalignment between CEFR-oriented writing descriptors and structurally grounded models of argumentation, as well as a notable lack of empirical work situated in Malaysian ESL contexts. Collectively, these findings suggest that existing assessment practices often prioritise practicality over explanatory depth, limiting their capacity to capture the cognitive and rhetorical dimensions of argumentative writing. Recent literature similarly cautions that without structurally explicit evaluation frameworks, assessments risk conflating language proficiency with argumentative competence, thereby obscuring learners' actual reasoning abilities (Wingate, 2021; Liu & Stapleton, 2023; Morris et al., 2024). This synthesis underscores the need for theoretically grounded approaches that can bridge linguistic performance and argumentative reasoning, providing a clearer foundation for both research and pedagogy in ESL writing assessment.

## Types of Argumentative Strength Evaluation

**Table 2 Classification of Argumentative Strength Evaluation Frameworks in ESL Writing**

| Category / Framework      | Description                                                                                                               |
|---------------------------|---------------------------------------------------------------------------------------------------------------------------|
| Analytic-model frameworks | Toulmin-based models; pragma-dialectical models   Explicit identification of claims, evidence, warrants, rebuttals   High |
| Rubric-based frameworks   | CEFR-aligned rubrics; IELTS writing descriptors   Holistic scoring; argumentation embedded in general criteria   Moderate |
| Computational frameworks  | Argument mining; AES systems   Automated feature extraction; indirect proxies for argument quality   Low–moderate         |

The review came up with three general categories of frameworks that are employed to assess argumentative strength in ESL writing: analytic-model frameworks, rubric-based frameworks and computational or evidence-based frameworks. The analytic-model frameworks are also based on the argumentation theory, and the internal structure of the arguments is explicitly represented. Of them, the Model of Argumentation proposed by Toulmin has become the most commonly used and referenced model especially when exploring the quality of reasoning, justification, and counter-argumentation (Stapleton and Wu, 2020; Qin and Karabacak, 2021). Their studies normally code the part of an argument like claims, data, warrants, and rebuttals, and through this, it is possible to analyze argumentative strength in a fine-grained way. Conversely, rubric-based systems are usually congruent with standardised writing tests and CEFR-informed definitions, in which argumentation is entrenched in more extensive standards of coherence, idea development, or task response (Ahmad and Kabilan, 2021; Uri, 2021). Although these rubrics are effective and scalable, they tend to be not specific to argument structure. Recently, computational and evidence-based models have been developed, which use automated scoring of the essay or argument-mining methods to identify the argumentative features using linguistic proxies (Balkir et al., 2023; Ke and Ng, 2024). These methods are more scaled, objective, but because of the use of surface-level indicators, there is concern regarding interpretability and construct validity especially when it comes to writing by ESL learners. Beyond differences in theoretical orientation, the reviewed frameworks also vary substantially in their pedagogical transparency and analytical applicability within ESL writing research. Analytic-model frameworks offer a clear advantage in making the logic of argumentation visible, thereby allowing researchers and educators to distinguish between linguistic adequacy and reasoning quality. Several recent studies emphasise that this explicit modelling of argumentative components supports more meaningful diagnostic feedback and facilitates alignment between instruction and assessment, particularly in academic and examination-oriented contexts (Wingate, 2021; Liu & Stapleton, 2023). In contrast, rubric-based frameworks, while effective for standardised assessment, often prioritise scoring efficiency over analytical depth, which may obscure learners' strengths and weaknesses in reasoning. This limitation has been noted in CEFR-aligned contexts, where broad performance descriptors can mask variation in argumentative competence among learners with similar proficiency levels (Hajar & Mahadi, 2022). Although computational and evidence-based approaches address concerns of scalability and objectivity, recent literature cautions that their reliance on proxy indicators limits their pedagogical usefulness, as such models offer little insight into how arguments are constructed or how learners might improve their reasoning (Shen & Yao, 2022; Ke & Ng, 2024). Collectively, these contrasts highlight the need for evaluation frameworks that balance analytical accuracy with contextual relevance, reinforcing the value of structurally explicit models for advancing both research and practice in ESL argumentative writing assessment.

Thorough comparison portrays distinct priorities in shaping how argumentative strength is interpreted and evaluated in each category of framework. Analytical-model frameworks emphasise theoretical precision and explicitly represent the internal structure of reasoning, making them particularly valuable for identifying strengths and weaknesses in students' argument construction. In contrast, rubric-based frameworks prioritise practicality and scalability, aligning closely with institutional assessment demands while often lacking the granularity needed to capture the quality of reasoning. Computational approaches offer competence and objectivity and rely heavily on linguistic proxies that may not fully represent the complexity of argumentative processes particularly in ESL contexts. From a pedagogical point of view, analytical-model frameworks appear most suitable for classroom use,

as they provide clearer diagnostic insights that can guide instruction and support the development of students' critical thinking and reasoning skills.

### Operationalisation of Argumentative Strength

**Table 3 Operationalisation of “Argumentative Strength” Across Framework Types**

| Category / Framework | Description                                                                                     |
|----------------------|-------------------------------------------------------------------------------------------------|
| Analytic-model       | Structural quality of reasoning   Presence and integration of claims, data, warrants, rebuttals |
| Rubric-based         | Overall effectiveness of argument   Coherence, clarity of ideas, task response                  |
| Computational        | Quantifiable argument features   Claim frequency, discourse markers, support relations          |

Building on the framework distinction outlined above, this section examines how analytic-model rubric-based and computational approaches operationalise argumentative strength. There was a significant difference in the operationalization of argumentative strength in the reviewed studies. Analytic-model frameworks theorise the strength of arguments as a quality, coherence, and integration of the elements of the argument, which focuses on logical connections among the claims, pieces of evidence, and reasoning (Qin and Stapleton, 2021; Liu and Stapleton, 2023). Strength is usually operationalised in these studies by presence of components, elaboration and functional congruency. In comparison, rubrics-based models are more likely to implicitly operationalise the strength of the argument by holistic terms, e.g. clarity of ideas, effectiveness of the argument, but does not explain how reasoning is graded (Hajar and Mahadi, 2022). This is because this absence of specific operational standards restricts diagnostic performance and creates confusion between linguistic fluency and argumentative thinking. The computational approaches indirectly operationalise the strength of argumentation by using quantifiable measures, which are either the frequency of claims, discourse markers, or syntactic complexity based on the assumption that these indicators are correlated with the quality of reasoning (Shen and Yao, 2022; Ke and Ng, 2024). Nevertheless, a number of studies warn that the effect of such proxies, especially in the ESL settings where linguistic limitations affect the distribution of features, might not be able to reliably measure the effectiveness of arguments. In general, the results show that there is no conceptual agreement, and it is necessary to develop structurally explicit models that can unite linguistic resources with argumentative arguments.

Besides conceptual inconsistency, another finding of the reviewed articles is a significant amount of variability between the level of transparency and replicability of various operationalisation strategies. Analytic-model frameworks usually utilize explicit coding schemes containing well-defined criteria of recognizing and assessing argument constituents, which contributes to the increased methodological reliability and allows conducting meaningful cross-study comparison (Qin and Karabacak, 2021; Liu and Stapleton, 2023). This transparency enables researchers to follow the contribution of particular linguistic resources to particular functions within an argument to support more subtle interpretation of argumentative performance. Conversely, rubric-based methods tend to have little methodological specificity on how evaluators differentiate between weak and strong arguments, which makes the methods less reliable and limits the applicability of results to instructional intervention (Hajar and Mahadi, 2022; Morris et al., 2024). On the same note, computational models, being methodologically the same in terms of algorithmic steps, provide minimal information about the cognitive mechanisms of constructing an argument, so it is challenging to match automated results with pedagogical aims (Shen and Yao, 2022; Ke and Ng, 2024). All these restrictions imply that in order to effectively operationalise argumentative strength, theoretical based frameworks are necessary which are not merely transparent and interpretable in order to facilitate systematic correlation of linguistic form, argumentative function and quality of reasoning in ESL writing.

Synthesising across the reviewed studies, three recurring dimensions of argumentative strength can be identified; a. structural completeness, referring to the presence and integration of core argument components; b. epistemic support reflecting the quality and relevance of evidence and justification; and c. dialogic engagement encompassing counter argumentation and rebuttal handling. To encapsulate, these three dimensions acts as a foundation for a more integrated conceptualization of argumentative strength that moves beyond fragmented and implicit definitions. Structural completeness ensures that arguments contain the necessary components while epistemic support evaluates the quality and relevance of justification. Dialogic engagement captures the ability to

address alternative perspectives and counterarguments. Collectively these dimensions offer a comprehensive framework for analysing how arguments are constructed and evaluated in ESL writing. Thus, an integrated perspective is particularly valuable in differentiating linguistic performance and reasoning ability while enabling a more precise understanding of learners' argumentative competence. It also provides a strong conceptual foundation for future research aiming to connect linguistic features, such as lexical complexity to specific argumentative functions. While analytical-model frameworks tends to operationalize all three dimensions explicitly, rubric based and computational approaches typically addresses these dimensions indirectly or partially contributing to conceptual inconsistency across studies.

#### Methodological Approaches

**Table 4 Methodological Strengths and Limitations of Evaluation Frameworks**

| Category / Framework | Description                                                                                                  |
|----------------------|--------------------------------------------------------------------------------------------------------------|
| Analytic-model       | High transparency; strong theoretical grounding; reliable coding   Labour-intensive; requires rater training |
| Rubric-based         | Efficient; scalable for examinations   Lacks structural detail; weak diagnostic value                        |
| Computational        | Scalable; objective scoring   Low interpretability; weak contextual adaptability                             |

Moving beyond classification, this synthesis reveals how the choice of evaluative framework shapes methodological design, data analysis and reporting practices. The studies reviewed shows a significant level of differences in research design, data analysis and reporting. Analytic-model based studies also tend to describe systematic coding methods and inter-rater reliability data, which improve the level of transparency and methodological trustworthiness (Stapleton and Wu, 2020; Qin and Karabacak, 2021). These practices make it possible to repeat the practices and enable cross-study comparison. Conversely, the studies that apply the rubric framework tend to focus on efficiency in assessment and correspondence with institutional grading processes, but offer minimal empirical information about the ways in which argumentative quality is differentiated as compared to overall writing competence (Uri, 2021; Morris et al., 2024). Computational studies usually provide algorithmic performance indicators, but only provide limited clarification of how extracted features are associated with theoretical constructs of argumentation (Balkir et al., 2023). Some of the authors bring up the issue of contextual validity, stating that the models, which are trained on the L1 or mixed corpora, may not be indicative of the argumentative practices of ESL learners (Ke and Ng, 2024). All these methodological differences will lead to inconsistent results and reduce the generalizability of results to educational settings.

Besides the differences in the methods of analysis, the reviewed studies also differ in their correspondence between the research methodology and the purpose of the pedagogy. Analytic-model research frequently follows a design that directly connects the evaluation criteria with the instructional goals, which allows a more coherent understanding of the learning process of argumentative skills and the possible factors that facilitate their formation with the use of specific feedback (Wingate, 2021; Liu and Stapleton, 2023). In comparison, rubric-oriented research often pays more attention to the results of summative assessment, which restricts its ability to influence instructional decision-making or formative assessment (Uri, 2021; Morris et al., 2024). Computational methods, despite being theoretically advanced, are mostly in touch with classroom realities since their results cannot be easily interpreted by teachers or learners and have little to guide them on how to improve argumentative thinking (Shen and Yao, 2022; Ke and Ng, 2024). A number of researchers claim that methodological strength in ESL writing studies ought to promote past statistical or algorithmic precision to interpretability, teaching applicability, and cultural sensitiveness (Hajar and Mahadi, 2022). These factors imply that methodological treatments of determining argumentative strength need to be able to balance analytical rigour with educational applicability so that the results may yield both theoretically significant and practically applicable results. Collectively, these methodological patterns indicate that analytic-model frameworks prioritise explanatory validity, rubric-based approaches prioritise assessment efficiency, and computational models prioritise scalability often at the expense of pedagogical interpretability.

#### Limitations and Contextual Gaps

**Table 5 Identified Research Gaps in ESL Argumentative Writing Assessment**

| Category / Framework  | Description                                              |
|-----------------------|----------------------------------------------------------|
| CEFR alignment        | Descriptors lack explicit argument structure             |
| Malaysian ESL context | Very limited empirical studies                           |
| Structural analysis   | Overreliance on holistic rubrics                         |
| Pedagogical usability | Computational models unsuitable for classroom assessment |

Across the reviewed studies, concerns regarding CEFR aligned writing assessment emerged primarily through reported limitations in rubric design and assessment transparency rather than as direct critiques of the CEFR framework itself. Several studies explicitly note difficulties in mapping CEFR writing descriptors onto the structural evaluation of argumentative reasoning particularly in relation to claims, justification, and counter argumentation (Ahmad & Kabilan, 2021; Hajar & Mahadi, 2022; Uri, 2021). Although CEFR promotes the ability to provide opinions and arguments in the higher levels of proficiency, it does not provide much advice on how to evaluate the quality of the reasoning process per se (Ahmad and Kabilan, 2021; Hajar and Mahadi, 2022). This lacuna has caused the excessive use of wholesome rubrics that do not effectively understand the structure of the argument. Moreover, the review shows that there is a lack of studies that are located in the Malaysian ESL setting, although nationwide curricula based on CEFR are applied. Majority of the empirical studies are focused on East Asian, Middle Eastern, or European, which restricts the applicability to the Malaysian classrooms and assessments (Uri, 2021; Thiagarajan and Hua, 2023). Moreover, not many studies analyze the interplay of language knowledge and argument structure directly, which creates a gap of little information on the developmental patterns of ESL argumentation. These existing gaps demonstrate the necessity of context-specific, theory-based studies that use clear analytic models to promote the body of knowledge concerning argumentative prowess in ESL writing.

Along with the problem of alignment and contextual coverage, other limitations associated with the reviewed literature demonstrate the problem of analytical focus and developmental perspective of the research of ESL argumentative writing. Cross-sectional designs, which record the argumentative performance at one moment, are adopted in many studies, which provides limited information about the development of argumentative strength over different levels of proficiency or educational stages (Wingate, 2021; Liu and Stapleton, 2023). This failure to have longitudinal or developmental analysis limits cognizance of how learners eventually combine linguistic material and argumentative reasoning. Moreover, the number of studies that explicitly investigate the role of particular linguistic attributes, e.g. lexical sophistication, grammatical complexity, or discourse organisation, in particular, argumentation elements, leads to disjointed conclusions about the connection between language proficiency and quality of reasoning (Shen and Yao, 2022). As scholars have observed, it is still hard to separate the issues with weak argumentative writing between the poor ability to reason and a lack of language expression without structurally explicit analytical frameworks (Hajar and Mahadi, 2022; Morris et al., 2024). These gaps provide significant reasons to consider the methodologically sound and theoretically based research designs which could possibly represent the structural as well as the developmental aspects of argumentative competence, particularly in underrepresented ESL contexts like Malaysia.

## DISCUSSION

This systematic review aims to synthesis how argumentative strength is conceptualized, operationalised and evaluated in ESL writing research with particular attention to the alignment between evaluative frameworks and underlying constructs of reasoning. By examining analytic-model, rubric- based and computational approaches side by side, the review moves beyond cataloguing assessment practices to reveal persistent conceptual and methodological tensions in the field (Wingate, 2021; Liu & Stapleton, 2023). Specifically, the synthesis highlights a recurring imbalance between the complexity of argumentative reasoning expected in academic writing and the evaluative mechanisms commonly used to assess it, particularly in CEFR- aligned contexts. The discussion that follows interprets these patterns in relation to construct validity, assessment purpose and contextual applicability, and considers their implications for future research and practice in ESL writing assessment.

By contrast, rubric-driven models, especially those that conform to CEFR writing descriptions, make an implicit operationalisation of quality in argumentation to general criterion, like task accomplishment or coherence. Although the above-mentioned rubrics contribute to standardisation in mass testing, they do not offer much help

in assessing the quality of reasoning and can establish a confusion between the skill of argumentation and the proficiency in the overall language (Ahmad and Kabilan, 2021; Hajar and Mahadi, 2022). It is this limitation that is most apparent in examination-based ESL settings, such as Malaysia, where argumentative writing is not directly judged by the structural criterion but is a requirement (Uri, 2021).

Computerized methods used in the assessment of argumentative writing have become more and more popular because they are scalable, efficient, and their objectivity can be achieved. Inferential argumentative strength based on linguistic proxies is the basis of automated essay scoring systems and argument-mining techniques which generally use discourse markers, frequency of claims, syntactic complexity or lexical properties to infer argumentative strength. But, as the results of this review demonstrate, there are still some issues concerning the interpretability and contextual validity of these methods, especially in ESL classes. Even linguistic proxies are not always reliable to predict the quality of reasoning, since the restricted resources of the language of ESL learners can limit their capacity to communicate their arguments, although there may be reasoning behind such arguments (Shen and Yao, 2022). Additionally, computational models have been shown to be trained on L1 or mixed-language data, which provokes concerns about their appropriateness to the study of the rhetorical strategies of ESL learners (Balkir et al., 2023; Ke and Ng, 2024). These drawbacks imply that computational tools are not enough to be applied as a substitute to human assessment, but rather they are inadequate as independent methods of assessing argumentative power in education studies.

Overall, the findings underscore the need for theoretically grounded and structurally explicit evaluation frameworks capable of distinguishing argumentative reasoning from surface linguistic features. Analytic-model approaches are particularly well suited to meeting this need, as they enable systematic analysis of reasoning structure while remaining adaptable to different proficiency levels and instructional contexts. In CEFR-oriented ESL settings where argumentation is expected but not clearly operationalized, such frameworks offer a principled means of bridging curricular expectations and assessment practice. By foregrounding the internal structure of arguments, analytic-model frameworks provide a robust foundation for advancing research on argumentative strength and for informing more transparent, valid, and pedagogically meaningful assessment practices in ESL writing.

## CONCLUSION

This systematic review examined frameworks used to evaluate argumentative strength in ESL student writing, synthesising findings across trends, framework types, operational definitions, methodological approaches, and contextual limitations. The review demonstrates that while a variety of assessment approaches exist, analytic-model frameworks, particularly those grounded in Toulmin's Model of Argumentation provide the most explicit and theoretically coherent means of evaluating argumentative reasoning. Rubric-based frameworks, despite their widespread use, often lack sufficient granularity to capture internal argument structure, while computational approaches face challenges related to interpretability and contextual adaptability. The review also reveals a persistent misalignment between CEFR writing descriptors and structural models of argumentation, as well as a lack of research situated in Malaysian ESL contexts. Despite its contributions, this review is limited to peer-reviewed studies published between 2010 and 2024 and primarily focuses on English-language publications. Consequently, relevant studies outside these criteria may not have been included.

Overall, the findings point to a clear need for structured, theory-driven approaches to argumentative writing assessment. Toulmin-based frameworks, especially when adapted to educational contexts, offer a robust foundation for future research examining the relationship between linguistic resources and argumentative performance in ESL writing

## REFERENCES

1. Ahmad, S. Z., & Kabilan, M. K. (2021). English language teachers' perceptions of the
2. CEFR-aligned curriculum in Malaysia. *Asian EFL Journal*, 28(1), 7–28.
3. Balkir, N., Ozturk, Y., & Aydin, Z. (2023). Automated assessment of argumentative
4. writing: A review of argument mining approaches in educational contexts. *Computer Assisted Language Learning*, 36(5–6), 934–957.
5. Booth, A., Sutton, A., & Papaioannou, D. (2021). *Systematic approaches to a successful*
6. *literature review* (3rd ed.). Sage.
7. Donato, R., & McCormick, D. (2021). Research synthesis in applied linguistics:
8. Methodological issues and directions. *Annual Review of Applied Linguistics*, 41, 1–17.
9. Gusenbauer, M., & Haddaway, N. R. (2020). Which academic search systems are suitable

- 
10. for systematic reviews or meta-analyses? *Research Synthesis Methods*, 11(2), 181–217.
  11. Hajar, S., & Mahadi, T. S. T. (2022). Teachers' perceptions of CEFR-aligned writing assessment in Malaysian secondary schools. *GEMA Online® Journal of Language Studies*, 22(2), 134–150.
  12. Heng, C. S., Tan, H., & Yusoff, N. M. (2023). Argumentative writing and higher-order thinking skills in ESL classrooms. *Journal of English for Academic Purposes*, 61, 101185.
  13. Ke, Z., & Ng, V. (2024). Argument mining in student essays: Challenges for second language writing assessment. *Language Learning & Technology*, 28(1), 1–22.
  14. Lee, I., & Deakin, L. (2016). Assessment of L2 writing: Teachers' perspectives and practices. *Assessing Writing*, 27, 1–13.
  15. Liu, F., & Stapleton, P. (2023). Argument structure and development in L2 argumentative writing. *Journal of Second Language Writing*, 59, 100963.
  16. Morris, K., Chan, S. H., & Khoo, S. M. (2024). Writing assessment practices and argumentative competence in ESL contexts. *Assessing Writing*, 60, 100708.
  17. Paul, J., & Criado, A. R. (2020). The art of writing literature review: What do we know and what do we need to know? *International Business Review*, 29(4), 101717.
  18. Qin, J., & Karabacak, E. (2021). Toulmin's model in L2 argumentative writing: Coding schemes and reliability issues. *System*, 98, 102482.
  19. Shen, X., & Yao, X. (2022). Automated evaluation of argumentative writing: Validity issues in ESL contexts. *Computer Assisted Language Learning*, 35(8), 1855–1877.
  20. Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. *Journal of Business Research*, 104, 333–339.
  21. Stapleton, P., & Wu, Y. (2020). Assessing argument structure in L2 writing: A Toulmin-based approach. *Journal of English for Academic Purposes*, 44, 100832.
  22. Thiagarajan, T., & Hua, T. K. (2023). CEFR-aligned ESL writing assessment in Malaysia: Challenges and implications. *3L: The Southeast Asian Journal of English Language Studies*, 29(1), 45–60.
  23. Uri, N. F. M. (2021). Examining CEFR-aligned writing rubrics in Malaysian ESL classrooms. *Malaysian Journal of Learning and Instruction*, 18(2), 95–116.
  24. Wingate, U. (2021). Argumentation in academic writing: Theory, pedagogy, and assessment. *Journal of English for Academic Purposes*, 49, 100944.
  25. Xiao, Y., & Watson, M. (2020). Guidance on conducting a systematic literature review. *Journal of Planning Education and Research*, 40(1), 93–112.
  26. Xiao, Y., Watson, M., & Papaioannou, D. (2021). Methodological rigor in systematic literature reviews in education. *Educational Research Review*, 34, 100401.