Check for updates

REVIEW

The measurement of text quality: current methods and open challenges

[version 1; peer review: 1 approved, 1 approved with reservations, 1 not approved]

Zuzana Toth [1,2]

[1]University of Modena and Reggio Emilia, Modena, Emilia-Romagna, Italy
[2]University of Basel, Basel, Basel, Switzerland

v1 **First published:** 07 Apr 2025, **5**:98
https://doi.org/10.12688/openreseurope.19312.1
Latest published: 07 Apr 2025, **5**:98
https://doi.org/10.12688/openreseurope.19312.1

Abstract

There is a general agreement that writing skills are essential in various areas of life, such as academic and professional achievement, lifelong learning, participation in society, etc (Valtin *et al.*, 2016). However, there is limited understanding of what constitutes a well-written text, how writing can be evaluated, to what extent the criteria that guide the evaluation can be articulated and standardised. This lack of clarity can place a significant burden on educators, who bear the responsibility of evaluating students' texts and provide constructive feedback. In some cases, it may also prompt concerns about the validity of high-stakes assessments that rely on the evaluation of written texts (Kobrin *et al.*, 2011). The present study aims to explore how writing quality is operationalised in research and how scholars ensure the validity and reliability of its measurement. It is based on the analysis of 60 studies published in two leading journals focusing on research on writing, *Assessing Writing* and *Journal of Second Language Writing*. The analyses show that the most established way to operationalise the construct of writing quality is by means of holistic or analytic evaluation scales. The reliability of the measurements is in most cases ensured by rater training and the calculation of inter-rater reliability or agreement; intra-rater reliability is rarely controlled for. The validity of the instruments is in most cases assumed rather than critically examined or discussed.

Keywords

text quality, writing quality, written proficiency, text assessment

Open Peer Review

Approval Status ? ✓ ✗

	1	2	3
version 1 07 Apr 2025	? view	✓ view	✗ view

1. **Heqiao Wang**, Michigan State University, East Lansing, USA
2. **Haticetül Kübra Er**, Erzurum Technical University, School of Foreign Languages, Erzurum, Turkey
3. **Jennifer Balade**, University of La Laguna, San Cristóbal de La Laguna, Spain

Any reports and responses or comments on the article can be found at the end of the article.

This article is included in the Horizon Europe gateway.

This article is included in the Marie-Sklodowska-Curie Actions (MSCA) gateway.

Corresponding author: Zuzana Toth (zuzana.toth@unibas.ch)

Author roles: Toth Z: Conceptualization, Formal Analysis, Funding Acquisition, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This project has received funding from the European Union's Horizon Europe research and innovation programme (Marie Skłodowska-Curie Actions, Grant agreement ID: 101066322)
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2025 Toth Z. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Toth Z. **The measurement of text quality: current methods and open challenges [version 1; peer review: 1 approved, 1 approved with reservations, 1 not approved]** Open Research Europe 2025, **5**:98
https://doi.org/10.12688/openreseurope.19312.1

First published: 07 Apr 2025, **5**:98 https://doi.org/10.12688/openreseurope.19312.1

Introduction

Strong writing skills are widely recognised as essential across various domains of life. Both the scientific literature (e.g., Bereiter & Scardamalia, 2013; Flower & Hayes, 1981; MacArthur *et al.*, 2019) and key framework and policy documents (e.g., Valtin *et al.*, 2016; The National Commission on Writing, 2003) highlight their pivotal role in academic and professional achievement, participation in society, lifelong learning, critical thinking, and the articulation of complex ideas. Essay writing is in fact a central component of various high-stakes assessments, such as language proficiency exams, final exams at the end of secondary education and college admissions tests. Although the evaluation of writing permeates multiple domains of life, there is limited knowledge of how much it is influenced by subjective factors and idiosyncrasies and to what extent it can be done in an objective and standardised way. It also remains unclear to what extent the criteria guiding the evaluation of texts can be explicitly articulated and to what extent the evaluation relies on intuitions honed through extensive professional experience (see Leki, 1995; Lines, 2012; Li, 1996).

Several studies suggest that the evaluation of writing is necessarily subject to significant individual and contextual variation. Tan *et al.* (2022), for instance, point out that writing is a complex, socially situated performance and the criteria for its assessment vary in relation to the socio-cultural context in which the communication is embedded, conventions related to disciplines, text types and text genres, etc. Kaplan (1966) describes a series of culturally influenced thought patterns that shape the construction of a paragraph across different languages and suggests that a text may be perceived as more or less coherent as a function of its adherence to the thought patterns expected by the reader, which may differ from those envisioned by the writer. Research on rater cognition (Connor-Linton, 1995; Cumming *et al.*, 2002; Kobayashi & Rinnert, 1996) also suggest that the perception of quality is likely to be influenced by the characteristics of the rater, such as his/her cultural or disciplinary background. This is not surprising, given that a text is a complex product and its evaluation, especially when made by a highly trained human, is informed by a rich set of criteria and arises from the interplay of a wide range of variables. The relative importance of these variables is, at least to a certain extent, subject to individual variation. This view is substantiated by research showing that language teachers, who engage in the evaluation of texts as a regular part of their professional practice, are often doubtful about the objectivity and adequacy of their own evaluations (Lines, 2012). Their doubts primarily stem from fluctuations in their own perceptions of text quality, discrepancies between their evaluations and those of their colleagues, and the belief that such variations render the evaluation "a matter of personal taste" (Lines, 2012).

A widely recognised approach to mitigating individual variation in text evaluation is the use of rating scales (also referred to as scoring rubrics). Scoring rubrics serve two main purposes: they guide evaluators to consider all relevant criteria, in some cases they also guide the raters in assigning a certain weight to certain aspects, and foster a shared understanding of what constitutes good writing. However, research indicates that raters often disagree on their scores and achieving an acceptable level of inter-rater agreement typically requires extensive training and calibration (Jonsson & Svingby, 2007; Schoonen, 2005). This practice however raises the question to what extent such scores reflect the rater's genuine perception of text quality rather than their ability to consistently apply a rubric. In other word, scoring rubrics may facilitate the achievement of high (or at least acceptable) inter-rater reliability at the expenses of validity.

Another problematic aspect is the relationship between the constructs of text quality and written proficiency. The two concepts are interrelated: the production of a well-written text presupposes high writing proficiency; vice versa, linguistic proficiency is largely judged on the basis of the quality of a linguistic performance such as writing. Nevertheless, the relationship between the two concept, quality and proficiency, remains underexplored and seldom critically examined in the literature. In fact, several studies examined here use the two terms interchangeably, to refer to the same construct (see DeVore & Kyle, 2023; Kim & Ro, 2024; Mostafa & Crossley, 2020).

The present study explores how researcher tackle the challenges in the measurement of text quality. It is guided by three research questions:

1. How is text quality measured in research?

2. What dimensions of writing are addressed in the measurement instruments?

3. How do researchers ensure the validity and the reliability of the measurement?

To address these questions, a selection of 60 papers was analysed. The articles were published between 2020 and 2024 in the two most influential journals on writing, as identified by Dong *et al.* (2024) through bibliometric research: *Assessing Writing* and *Journal of Second Language Writing*. The period of examination was restricted to four years to maintain a manageable dataset. Initially, a broader search using the keywords *writing quality*, *text quality and written/writing proficiency* yielded 134 articles, which were downloaded and manually screened, to identify those most relevant to our research focus. Among papers focusing on written proficiency, we selected those that appeared to treat the two concepts, proficiency and quality, as broadly referring to the same construct. For instance, Thwaites *et al.* (2024), who investigate proficiency ratings by means of comparative judgment, use the terms proficiency and quality as interchangeable in various points of their article, as exemplified by the following quote: "Regarding proficiency span, conducting comparisons is easiest where items differ widely in proficiency/quality, since it is easiest to decide which of two texts is "better" when they occupy different locations on the proficiency spectrum" (p. 3). In a similar vein, DeVore and Kyle (2023) use the terms "writing quality scores" and

"writing proficiency scores" as synonyms.[1] In addition, two studies, conducted by Bulté & Housen (2014) and by MacArthur *et al.* (2019) were included because they are inherent to the focus of the present study and are listed among the most influential articles by Dong *et al.* (2024). Studies focusing on the quality of integrated writing were excluded, as evaluating such performances poses unique challenges (e.g., how to assess the degree of textual borrowing?).

While we acknowledge that limiting the main scope of the analysis to two journals and a four-year period provides only a partial perspective on the research, we believe it encompasses most articles that have significantly influenced the field in the past four years. The following sections are organised around the three research questions.

1. How is text quality measured in research?

Written texts, like complex performances in general, can be evaluated based on a broad range of criteria. One of the main challenges in ensuring an objective and fair evaluation is to guarantee that all relevant criteria are given the necessary consideration and that spurious criteria don't unduly influence the evaluation. Additionally, it is important to consider whether all relevant criteria should be weighted equally or if a hierarchical weighting is necessary, for example by assigning greater importance to higher level aspect such as content and coherence than to lower-level aspects such as grammatical correctness and orthography.

The most widely established method for the evaluation of writing is the use of evaluation scales, also referred to as scoring rubrics. As Weigle (2002, p. 112) points out, the use of scoring rubrics was proposed as a response to the shortcomings of an earlier assessment method known as general impression marking, which consists in assigning a score to the texts without explicitly stating the criteria that are considered. One of the greatest benefits of scoring rubrics is the transparency and explicitness of the criteria considered by the evaluators (Frederiksen & Collins, 1989; Jonsson & Svingby, 2007), which also contributes to a shared understanding of what is considered to be 'good', what evaluators look for and value. In instructional settings, such shared understanding may promote motivation, metacognitive skills and ultimately have a positive impact on the quality of the students' performance (Bissell & Lemons, 2006; He *et al.*, 2022; Wolf & Stevens, 2007).

According to Weigle (2002, p. 109), three types of writing scales can be distinguished: primary trait scales, holistic scales and analytic scales. Primary trait scales are tailored to a specific writing task and reflect its unique demands. As these scales are task specific and highly detailed, their development and application is time intensive. Consequently, primary trait scales are rarely used in practice, even though they provide rich information about a person's writing ability, especially if a variety of texts is collected (Weigle, 2002, p. 110). Holistic scales evaluate a text by assigning it to a specific performance level, with each level providing a detailed description of how the text performs across various dimensions, such as content, coherence, cohesion, language use, etc. These levels encapsulate multiple perspectives on the text, offering a well-rounded view of its overall quality within a single rating. This method preservers the transparency by offering clearly defined performance levels that articulate the key dimensions of writing quality, but is more time efficient than primary-trait scoring. In fact, holistic scales are commonly used in large-scale and high-stakes assessment, such as the language exams TOEFL and the college admission exam SAT. A notable limitation of holistic scales is their inability to capture potential discrepancies across different dimensions of text quality. For example, a text might exhibit well-organised and engaging content but at the same time contain numerous linguistic errors. In such a case, the text should awarded a higher score for content and coherence, but a lower score for language use. However, holistic scales require the rater to assign a single performance level encompassing all aspects, thereby obscuring these details. Analytic scales address this issue by enabling evaluation on separate subscales, each focusing on a specific aspect of the text. For instance, one of the most widely adopted analytic scales for assessing second language writing, proposed by Jacobs *et al.* (1981), decomposes writing quality into five distinct subscales: content, organization, vocabulary, language use, and mechanics. However, even though subscales allow for a more detailed evaluation of specific areas of strengths and weaknesses, a final judgment often requires the aggregation of the individual scores into a single score, blurring most of the details. The final score may be calculated using a weighted combination of the sub-scores, for instance, by assigning greater weight to content and organisation compared to language use and mechanics (see Connor-Linton & Polio, 2014). However, this approach raises concerns of validity, as the predefined weights assigned to features may not reflect what captures the reader's attention (e.g., as Colombo, 2001 notes, certain errors in language use may be more disruptive and distracting others). Consequently, the final score may not align with the reader's authentic perception of text quality.

Finally, a more innovative method to evaluate text quality is comparative judgments (CJ) (Pollitt, 2012). In this case, raters don't assign an absolute value to a text but instead repeatedly compare two texts and select the one they deem of higher quality. The series of judgments is subsequently transformed into a scale, where performances are ranked in order of increasing quality. This method has gained popularity over the past decade; however, several aspects require further research. In fact, comparative judgment may address some of the limitations associated with scoring rubrics, such as inter- and intra-rater reliability and validity, yet it also presents its own challenges. One of the key concerns is that similar to general impression

[1] The relationship between the concepts of text quality and written proficiency warrants careful consideration to determine whether these terms refer to the same construct, different facets of a single construct, or distinct constructs. However, this topic remains largely unexplored in the literature and its comprehensive evaluation lies beyond the scope of the present article.

marking this method does not allow for the explicit articulation of the criteria on which raters base their judgments.

In the studies analysed here, the instruments to measure text quality were distributed as follows: analytic scale in 22 studies (Table 1), holistic scale in 16 studies (Table 2), genre specific scale in four studies (Canz *et al.*, 2020; Deng *et al.*, 2022; Uzun, 2021; Zhang & Lu, 2022), which resemble the level of detail and specificity of the primary trait scales analysed by Weigle (2002). CJ was used only in two cases, in conjunction with scoring rubrics (Siekmann *et al.*, 2022; Thwaites *et al.*, 2024). Seven studies used a combination of different types of scales: both holistic and analytic scales (Fladung *et al.*, 2023; Hadidi, 2023; Imbler *et al.*, 2023; McCarthy *et al.*, 2022); two different scales respectively for content and basic aspects of writing (Gong *et al.*, 2022); a holistic scale and a simplified rating of communicative adequacy (Schmidgall & Powers, 2020); two different (but not explicitly genre specific) rubrics for the evaluation of narrative and argumentative texts (Abdi Tabari & Johnson, 2023); one study employed a specifically developed checklist rather than a scale (Effatpanah *et al.*, 2024). Nine studies did not report a sufficiently detailed descriptions of their text evaluation methods: two studies (Kim & Ro, 2024; Khushik, 2024) reported using a scale from the Common European Framework of Reference for Languages (CEFR), which encompasses both holistic and analytic scales, without indicating the specific scale employed; in one study, text quality was assumed based on the author's experience, as the researchers compared essays written by university students with papers published in leading scientific journals (Li & Yang, 2023); four studies used the grades assigned to the texts by a testing authority, without a detailed discussion of how these grades were awarded (Appel *et al.*, 2023; Chen, 2020; Lyashevskaya *et al.*, 2021; Li *et al.*, 2023).

Table 1 and Table 2 report the specific analytic and holistic scales encountered and their distribution across the studies. Research employing an analytic scale (Table 1) focuses almost exclusively on English texts: only two out of the 22 studies focus on languages other than English, specifically Greek (Emmanouil *et al.*, 2024) and German (Riemenschneider *et al.*, 2021). Both studies analyse texts written by secondary school students as part of their final certification assessments and evaluate them according to the guidelines established by the relevant educational authorities responsible for overseeing these examinations.

When looking at the 20 studies focusing on English texts, the ESL Composition Profile proposed by Jacobs *et al.* (1981) emerges as the most frequently used scale, employed in 10 cases (its original version in 8 cases, its adaptation by Connor-Linton and Polio (2014) in 2 cases). Three studies adopt scales developed by testing or educational authorities, while seven employ specifically designed scales. Among these seven studies, three

Table 1. Overview of studies employing an analytic scale.

Type of analytic scale	Studies
ESL composition profile (Jacobs *et al.*, 1981) or its adapted version by Connor-Linton & Polio (2014) [studies using the adapted version are marked with*]	Bulté & Housen, 2014* Chang, 2023* Eguchi & Kyle, 2023 Li & Yu, 2024 Peng *et al.*, 2023 Peng *et al.*, 2024 Shafiee Rad & Alipour, 2024 Teng & Zhang, 2020 Teng, 2022 Vandermeulen *et al.*, 2024
study-specific analytic rubric	Zhang & Chen, 2022 Nguyen *et al.*, 2024 Pun & Li, 2024 Wang & Fan, 2020 Rock, 2022 Wang & Xie, 2022
Global Test of English Communication scale	Yasuda, 2024
Pennsylvania Writing Assessment Domain Scoring Guide	Ghaffar *et al.*, 2020
General written production assessment guidelines of the Greek Ministry of Education	Emmanouil *et al.*, 2024
IELTS scale	Liu & Liu, 2023
Overview of the aspects of linguistic complexity covered in the German educational standards and recommended grading rubrics.	Riemenschneider *et al.*, 2021

Table 2. Overview of studies employing a holistic scale.

Type of holistic scale	Studies
TOEFL	Mostafa & Crossley, 2020 Qian, 2023 Vasylets & Marín, 2021
Chinese Proficiency Test (HSK)	DeVore & Kyle, 2023 Hao *et al.*, 2023; Hao *et al.*, 2024
CET-4 rubric (National College English Tests-Band Four in China)	Kim & Kessler, 2022 Li & Huang, 2022
CET-6 rubric (National College English Tests-Band Six in China)	Yang *et al.*, 2023
Scale developed by UCI Writing Project	Maamuujav, 2021 Maamuujav *et al.*, 2021
Study-specific holistic scale	Liao, 2023; MacArthur *et al.*, 2019
Gaokao English writing	Zhang & Ouyang, 2023
SAT	Kim *et al.*, 2021
Duolingo English Test (DET)	Naismith *et al.*, 2024

(Rock, 2022; Wang & Fan, 2020; Wang & Xie, 2022) focus on the development of the scale itself as the primary objective. Rock (2022) proposes a scale for the evaluation of academic blog posts; Wang and Fan (2020) focus on the development and validation of a scale for assessing business English writing, while Wang and Xie (2022) aim to design a scale for evaluating discourse competence.

Similar to the group of studies employing analytic scales, those using holistic scales (Table 2) predominantly focus on English, with four exceptions (DeVore & Kyle, 2023; Hao *et al.*, 2023; Hao *et al.*, 2024; Liao, 2023) that investigate writing in Chinese. Among the holistic rubrics, there is a slight preference for the TOEFL rubric, employed in four studies, three of which use an adapted version. The remaining studies either utilise evaluation scales developed by educational or testing authorities—such as the Chinese Proficiency Test rubric (three studies), the CET-4 rubric (two studies), the SAT rubric, the Gaokao rubric, the UCI Writing Project scale—or rely on specifically designed scales created for individual studies.

To summarise, researchers use a variety of tools to measure text quality. The most widely used tools are holistic and analytic rating scales; in contrast, genre specific scales and comparative judgment are less frequent. Among studies using holistic or analytic scales, there is a noticeable preference for those developed by testing authorities, likely for two main reasons. First, as these scales undergo a rigorous validation process, allowing researchers to rely on their established validity without the need for additional validation efforts. Second, several studies use the scores provided by the testing authorities (see Section 3), either as a sole source of scoring, or as a benchmark for calculating inter-rater reliability by comparing the authority's scores with those assigned by specifically recruited raters. This practice, however, raises some concerns, as scales designed for large-scale assessments may lack the level of detail and specificity needed to address the nuanced requirements of research. Moreover, the remarkable variety of instruments employed to measure text quality raises the question to which extent these tools converge in assessing a common underlying construct. In order to further investigate this question, the next section presents a qualitative analysis of the scales.

2. Which aspects of writing are addressed by the measurement instruments?

A qualitative content analysis employing a deductive approach (Mayring, 2014) was conducted to gain deeper insights into the aspects of writing addressed by the different holistic and analytic scales presented in Table 1 and Table 2. The analysis focuses on scales used to evaluate English texts, for two reasons: first, other languages are underrepresented in the studies; second, the scales used for assessing texts written in Chinese—the only language besides English investigated in more than one study—were not always described in sufficient detail in the studies or could not be retrieved online. Additionally, the analytic scale developed by Wang and Xie (2022) was excluded from this analysis, as it differs from the other scales by placing a fine-grained focus on a specific aspect of text quality—coherence and cohesion—rather than providing an assessment of overall text quality. Scales proposed by educational or testing authorities will be referred to by their official names or well-known acronyms (e.g., IELTS), while study-specific scales will be identified by referencing the respective study.

The analysis of analytic scales reveals the existence of five recurrent subscales: content, organization (also referred to as coherence and cohesion), language use, vocabulary, and mechanics, which can be seen as fundamental dimensions of text quality. Each analytic scale includes at least three of these dimensions. The absence of a specific subscale doesn't necessarily indicate that its underlying dimension is not evaluated; on the contrary, in most cases it has been subsumed within another subscale. In fact, these dimensions are complex and interrelated, often blending into one another, with no clear-cut boundaries separating them.

All scales include a subscale dedicated to the evaluation of the linguistic resources employed in writing. In some cases, this subscale has a broader scope and is referred to as *language use*, while in others it adopts a narrower focus and is labelled *grammar*. Regardless of the terminology, the consistent inclusion of a language-related subscale highlights the importance of this dimension as a fundamental component of text quality. The features assessed under this subscale include the degree of linguistic accuracy (often expressed in terms of error frequency), complexity and range of grammatical structures, as well as variety of sentence patterns (e.g. balanced used of simple, complex and compound sentences). In three cases (Jacobs *et al.*, 1981; Connor-Linton & Polio, 2014; IELTS), the subscale goes a step further, by also considering the impact of errors on comprehension or the severity of errors.

In six cases (Connor-Linton & Polio, 2014; Jacobs *et al.*, 1981; IELTS; Nguyen *et al.*, 2024; Wang & Fan, 2020), the language use subscale is distinct from the subscale dedicated to vocabulary, allowing for a more nuanced description of both aspects. In contrast, in four cases (Keller *et al.*, 2024; Pun & Li, 2024; Rock, 2022; Pennsylvania Writing Assessment Domain Scoring Guide), the language use subscale encompasses vocabulary, which testifies the interconnectedness of the two domains. For instance, when considering accuracy, errors in word formation, spelling and morphosyntax (e.g. agreement) cannot always be clearly distinguished from each other, making it more suitable to assess vocabulary within a broader scale of language use. On the other hand, however, research indicates that lexical diversity, a multidimensional construct itself, has a significant impact on human perception of text quality (Kim & Crossley, 2018; Kyle *et al.*, 2021; Woods *et al.*, 2023). This observation suggests that vocabulary should be evaluated as an independent component.

A subscale addressing content is present in seven scales, underscoring its significance as a key aspect of text quality. Although the IELTS scale does not feature a dedicated content subscale, content-related elements are integrated into the subscale labelled Task Response, which evaluates how effectively the writer addresses the task and develops ideas.

To illustrate this point, Table 3 provides a comparison of the descriptors of the highest level (level 4) of the subscale referred to as "Content" in the ESL Composition Profile (Jacobs *et al.*, 1981), its revised version by Connor-Linton & Polio (2014), and the corresponding highest level (band score 6) of "Task Response" in the IELTS scale. Each descriptor is quoted verbatim from the respective source documents. This comparison highlights how each scale conceptualises and prioritises content-related criteria, reflecting slightly different emphases on the development and relevance of ideas.

As illustrated in Table 3, the descriptors referring to the requirement of substantive content and throughout and logical development of the thesis (Content) closely align with those referring to clear and fully developed position and in depth-exploration of the prompt (Task Response). Similarly, all three subscales make reference to the relevance of the presented ideas.

Another highly salient aspect is Organisation, a subscale present in all but two scales. The two exceptions (IELTS, Wang & Fan, 2020), however, include a subscale called Coherence and cohesion. In fact, these two subscales–Organisation on one side and Coherence and cohesion on the other–largely address the same aspects, such as the internal architecture of paragraphs, the connection of ideas within and between paragraphs, the presence of an introduction and a conclusion, and the use of connectors to explicitly signal logical connections. The impression that the content and focus of these two subscales aligns significantly is further confirmed by the fact that no scale contains both subscales simultaneously. Thus, the difference appears to be terminological rather than conceptual. This observation is illustrated in Table 4, which reports the

Table 3. Overlapping descriptors in the subscales Content and Task response.

Scale	ESL Composition Profile (Jacobs *et al.*, 1981)	Scale by Connor-Linton & Polio (2014)	IELTS scale
Subscale	Content	Content	Task Response
Overlapping descriptors	substantive • thorough development of thesis	Thorough and logical development of thesis Substantive and detailed	The prompt is appropriately addressed and explored in depth. A clear and fully developed position is presented which directly answers the question/s.
	relevant to assigned topic	No irrelevant information	Ideas are relevant, fully extended and well supported.

Table 4. Overlapping descriptors in the subscales Organisation and Coherence & cohesion.

Scale	ESL Composition Profile (Jacobs *et al.*, 1981)	Scale by Connor-Linton & Polio (2014)	IELTS scale
Subscale	Organisation	Organisation	Coherence and cohesion
Overlapping descriptors	well-organized logical sequencing	Substantive introduction and conclusion Excellent connections between paragraphs Unity within every paragraph	Paragraphing is skilfully managed.
	cohesive	Excellent use of transition word	Cohesion is used in such a way that it very rarely attracts attention. Any lapses in coherence or cohesion are minimal.

highest band descriptors for Coherence and Cohesion in the IELTS scale (band 6) alongside the highest-level descriptors for organisation in the ESL Composition Profile and its revised version (band 4). It is worth nothing that, in addition to the similarities, the IELTS scale includes unique descriptors not found in other scales, referring to the effort needed to follow the message and the presence of lapses in coherence and cohesion.

Finally, a subscale dedicated to Mechanics is present in four cases (ESL Composition Profile, Connor-Linton & Polio, 2014; Rock, 2022, Pennsylvania Writing Assessment Domain Scoring Guide). This subscale addresses aspects such as mastery of conventions, spelling, and layout. In three cases, punctuation is included as part of this subscale, which is a controversial decision from a theoretical standpoint. In fact, studies in both theoretical and applied linguistics (Ferrari, 2014; Moore, 2016) show that interpunction plays a significant role in text cohesion and logical structuring. This evidence suggests that punctuation may warrant consideration within a broader textual coherence framework rather than being confined solely to Mechanics.

In addition to these common aspects, some of the analytic scales include additional dimensions, such as the presence of genre-specific features (Rock, 2022, Global Test of English Communication, Pennsylvania Writing Assessment Domain Scoring Guide), strategic competence and intercultural competence (Wang & Fan, 2020), business knowledge (Wang & Fan, 2020), style and the presence of a writer's voice (Global Test of English Communication, Pennsylvania Writing Assessment Domain Scoring Guide).

Among the holistic scales listed in Table 4, the analysis focuses on those specifically designed to measure writing quality in English and that were either reported in the studies using them or publicly accessible online. Based on these criteria, four scales were selected: STA, TOEFL, CET-4, and DET. Similar to the analytic scales, these holistic scales address five main aspects of writing: content, coherence and cohesion, language use, vocabulary, and accuracy. It is worth noting that the original CET-4 scale does not include a specific reference to vocabulary. However, Kim and Kessler (2022) opted to incorporate this dimension, as doing so improved inter-rater reliability. Similar to the IELTS scale, the TOEFL and the SAT scale also make reference to task fulfilment. The only major difference between the holistic and the analytic scales lies in the level of detail: holistic scales provide less granular descriptions, with all aspects integrated into a single band descriptor rather than being separately articulated.

The analyses presented here suggest that, despite terminological differences and some discrepancies related to the classification of features into subscales, most measurement instruments refer to similar core aspects of writing quality. This homogeneity points towards a common understanding of what are the main dimensions of the construct. The main differences between the various scales lie in two areas: 1) how the dimensions of writing quality are compartmentalised into subscales; 2) the level of detail provided in the description of each performance band. In other words, while there appears to be a shared understanding of the core dimensions of writing quality, articulating them in a way that fully captures the construct's nuanced nature proves problematic.

3. How do researchers ensure the validity and the reliability of the measurement?

Among the 60 studies analysed here, ten studies used the text scores assigned by testing authorities (Appel *et al.*, 2023; Chen, 2020; DeVore & Kyle, 2023; Eguchi & Kyle, 2023; Hao *et al.*, 2023; Li *et al.*, 2023; Lyashevskaya *et al.*, 2021; Qian, 2023; Schmidgall & Powers, 2020); one study didn't employ scoring, but divided the texts into two groups (low and high quality) based on the author's experience in academic writing (Li & Yang, 2023); three studies analysed texts that had been previously scored in earlier research (Bulté & Housen, 2014; Gong *et al.*, 2022; Uzun, 2021); two studies report the extensive experience of the raters as evidence supporting the reliability of their evaluations (Emmanouil *et al.*, 2024; Kim & Ro, 2024). In four cases, the rating scale

itself was the final product, developed collaboratively with the raters (Chen, 2020; Rock, 2022; Wang & Fan, 2020), while one study specifically investigated rater bias (Liu & Liu, 2023).

The remaining studies employ a wide range of strategies to ensure the reliability of text quality measurements. These include practices such rigorous rater training, scoring texts with the involvement of at least two raters and employing various statistical techniques to assess inter-rater agreement or reliability. Specifically, 38 studies report that the entire texts corpus or at least part of it was evaluated by at least two raters. The proportion of texts that underwent multiple scoring ranges from 7% to 100%, as shown in Table 5. Notably, the majority of the studies (29 out of 38) ensured that the entire dataset was scored by at multiple raters, reflecting a strong commitment to achieve reliable scoring.

Among the 38 studies that report scoring by multiple raters, 5 studies don't report on inter-rater agreement or reliability. In three cases (Mostafa & Crossley, 2020; Yang *et al.*, 2023; Zhang & Lu, 2022), the scoring procedure was designed to account for both raters' evaluations rather than assessing inter-rater agreement. Specifically, in the studies conducted by Mostafa & Crossley (2020) and Zhang & Lu (2022), the final score assigned to an essay was calculated as the average of the two raters' scores, provided the difference between their scores did not exceed one point. If the discrepancy was greater than one point, a third rater was brought in to resolve the difference. In Yang *et al.*'s (2023) study, the final score was the average of scores provided by three raters. In the remaining two cases (Peng *et al.*, 2024; Zhang & Chen, 2022), the authors indicate that the entire dataset was scored by multiple raters but do not provide any details on inter-rater agreement.

Table 6 presents an overview of the statistical procedures employed to measure inter-rater reliability or inter-rater agreement in the reviewed studies. A significant proportion of these studies evaluate inter-rater reliability through correlation analyses, with Pearson's correlation coefficient being the most employed metric. A few studies report r-values without specifying the statistical technique used, which may limit the interpretability and replicability of their findings. Several studies employ Kendall's rank-based measures (Kendall's Coefficient of Concordance, Kendall's τb, or Kendall's Rank Correlation) and the Intraclass Correlation Coefficient (ICC). A smaller subset of studies incorporates Rasch modelling to refine scoring procedure. These include double-rating adjustment through Rasch models or the application of Rasch-Kappa measures. Other advanced statistical approaches, such as the calculation of G- and Phi-coefficients, based on generalisability theory, and mixed-effect models with random intercepts, further showcase the diversity of methodological strategies aimed at ensuring reliability in rating processes.

Inter-rater agreement, on the other hand, is most frequently assessed using Cohen's Kappa, followed by the percentage of agreement, a straightforward but less sophisticated metrics. Krippendorff's Alpha is only used only in one study. Intra-rater agreement is only examined in two studies (Teng & Zhang, 2020; Wang & Xie, 2022), and none of the studies dedicates any detailed consideration to the validity of the instrument used to measure text quality.

Table 5. Studies with multiple text ratings.

Portion of texts scored by multiple raters	Reference to the study
(approximately 7%)	Effatpanah *et al.*, 2024
10%	Abdi Tabari & Johnson, 2023
18%	Nguyen *et al.*, 2024
20%	Deng *et al.*, 2022 Teng, 2022 Vandermeulen *et al.*, 2024
27%	Wang & Xie, 2022
30%	Vasylets & Marín, 2021
40%	Fladung *et al.*, 2023
50%	Li & Yu, 2024
100%	Canz *et al.*, 2020 Chang, 2023 Ghaffar *et al.*, 2020 Hadidi, 2023 Keller *et al.*, 2024 Kim *et al.*, 2021 Kim & Kessler, 2022 Li & Huang, 2022 Liao, 2023 MacArthur *et al.*, 2019 Maamuujav, 2021 Maamuujav *et al.*, 2021 McCarthy *et al.*, 2022 Mostafa & Crossley, 2020 Naismith *et al.*, 2024 Peng *et al.*, 2023 Peng *et al.*, 2024 Pun & Li, 2024 Rock, 2022 Shafiee Rad & Alipour, 2024 Siekmann *et al.*, 2022 Thwaites *et al.*, 2024 Woods *et al.*, 2023 Yang *et al.*, 2023 Yasuda, 2024 Zhang & Chen, 2022 Zhang & Lu, 2022 Zhang & Ouyang, 2023

4. Conclusions

The analyses presented here reveal that the most frequently used instruments to measure text quality in research are analytic rating scales, present in 26 studies (serving as the sole method of evaluation in 22 cases and used in combination with holistic scales in four cases). The most frequently used analytic scale is the one proposed by Jacobs *et al.* (1981) along with its adaptation by Connor-Linton & Polio (2014). The prevalence of analytic scales is likely tied to the focus on

Table 6. Statistical procedures to measure inter-rater reliability or inter-rater agreement.

Pearson correlation coefficient	Chang, 2023 Kim & Kessler, 2022 Peng *et al.*, 2023 Effatpanah *et al.*, 2024* Abdi Tabari & Johnson, 2023
Kendall's Coefficient of Concordance	Deng *et al.*, 2022
Kendall's τb	Fladung *et al.*, 2023
Kendall's Rank Correlation	Liao, 2023
Intraclass Correlation Coefficient (ICC)	Canz *et al.*, 2020* Li & Yu, 2024 McCarthy *et al.*, 2022* Vandermeulen *et al.*, 2024 Zhang & Ouyang, 2023
r- values (specific statistical technique not reported)	MacArthur *et al.*, 2019* Teng & Zhang, 2020 Teng, 2022 Wang & Xie, 2022
Double-rating with Rasch model adjustment	Khushik, 2024 Woods *et al.*, 2023
Rasch-Kappa	Thwaites *et al.*, 2024
G- and Phi-coefficient	Li & Huang, 2022
Mixed effect models with random intercepts	Riemenschneider *et al.*, 2021
Cohen's Kappa	Canz *et al.*, 2020* Effatpanah *et al.*, 2024* Imbler *et al.*, 2023 Keller *et al.*, 2024 Kim *et al.*, 2021 McCarthy *et al.*, 2022* Naismith *et al.*, 2024 Shafiee Rad & Alipour, 2024 Vasylets & Marín, 2021
Percentage of agreement	Ghaffar *et al.*, 2020 Maamuujav, 2021 Maamuujav *et al.*, 2021 MacArthur *et al.*, 2019* Nguyen *et al.*, 2024 Pun & Li, 2024
Krippendorff's Alpha	Yasuda, 2024

Note: studies marked with * are listed twice because they employ two different techniques

second language learners, who often exhibit uneven profiles across different aspects of writing quality (Weigle, 2002, p. 120).

The second most widespread method is the use of holistic scales, primarily those developed by testing authorities. Scales tailored to specific genres, which provide a level of detail comparable to the primary trait scales analysed by Weigle (2002, p. 110–112) and comparative judgment are used only in a limited number of studies.

Despite the abundance of analytic and holistic scales, there appears to be a general consensus on the core aspects of writing quality, as the various scales tend to evaluate the same aspects, i.e. content, language use, vocabulary, coherence and cohesion, mechanics. The main differences lie in the level of detail in which these aspects are treated in the band descriptors and in the way they are compartmentalised into subscales in analytic scales. Notably, these aspects are interrelated, with no clear-cut boundaries between them. The process of dividing them into distinct subscales remains insufficiently explored and warrants deeper investigation.

The appropriateness of rating scales as tools for measuring text quality is mostly assumed and seldom subjected to critical scrutiny. On the one hand, this can be justified by the rigorous and systematic validation process to which scales developed by testing authorities are subjected. On the other hand, a scale designed and validated for language testing may not be automatically transferable to research contexts. For instance, to deepen our understanding of human judgments of text quality are formed, it may be necessary to differentiate between range of linguistic structures and grammatical accuracy–elements that are usually merged into a single subscale dedicated to language use (Pallotti, 2009).

Further research is needed to explore the primary strengths and weaknesses of the various scales and to determine the extent to which these scales measure the same underlying construct. Moreover, the increasing interest in the use of comparative judgment calls for research on the concurrent validity of these two assessment methods.

The wide range of statistical procedures employed across the studies to ensure the reliability of the measurements underscores the need for future research to undertake a comprehensive evaluation of the strengths and limitations of these approaches. Such studies could provide critical insights into the suitability of each method for different contexts and help establish best practices—especially for linguists, who are not specialised in statistics—for ensuring reliable measurements.

Additionally, greater attention should be directed toward examining intra-rater variability, an often-overlooked aspect of rater consistency that could significantly impact the reliability of scoring.

Ethics and consent

Ethical approval and consent were not required.

Data availability

No data associated with this article.

Acknowledgment

The author would like to thank Gabriele Pallotti for his insightful feedback on an earlier version of this paper.

References

Abdi Tabari M, Johnson MD: **Exploring new insights into the role of cohesive devices in written academic genres.** *Assess Writ.* 2023; **57**: 100749.
Publisher Full Text

Appel R, Arias A, Baker B, *et al.*: **Insights from lexical and syntactic analyses of a French for academic purposes assessment.** *Assess Writ.* 2023; **58**: 100789.
Publisher Full Text

Bereiter C, Scardamalia M, (Eds.).: **The psychology of written composition (0 ed.).** Routledge, 2013.
Publisher Full Text

Bissell AN, Lemons PP: **A new method for assessing critical thinking in the classroom.** *BioScience.* 2006; **56**(1): 66–72.
Publisher Full Text

Bulté B, Housen A: **Conceptualizing and measuring short-term changes in L2 writing complexity.** *J Second Lang Writ.* 2014; **26**: 42–65.
Publisher Full Text

Canz T, Hoffmann L, Kania R: **Presentation-mode effects in large-scale writing assessments.** *Assess Writ.* 2020; **45**: 100470.
Publisher Full Text

Chang TS: **A move analysis of Chinese L2 student essays from the sociocognitive perspective: genres, languages, and writing quality.** *Assess Writ.* 2023; **57**: 100750.
Publisher Full Text

Chen Y: **Assessing the lexical richness of figurative expressions in Taiwanese EFL learners' writing.** *Assess Writ.* 2020; **43**: 100414.
Publisher Full Text

Colombo A: **La valutazione della prima prova dell'esame di stato.** In: A. R. Guerriero (Ed.), *Laboratorio di scrittura. Non solo temi all'esame di Stato. Idee per un curricolo.* La Nuova Italia, 2001; 253–271.
Reference Source

Connor-Linton J: **Crosscultural comparison of writing standards: American ESL and Japanese EFL.** *World English.* 1995; **14**(1): 99–115.
Publisher Full Text

Connor-Linton J, Polio C: **Comparing perspectives on L2 writing: multiple analyses of a common corpus.** *J Second Lang Writ.* 2014; **26**: 1–9.
Publisher Full Text

Cumming A, Kantor R, Powers DE: **Decision making while rating ESL/EFL writing tasks: a descriptive framework.** *Mod Lang J.* 2002; **86**(1): 67–96.
Publisher Full Text

Deng Z, Uccelli P, Snow C: **Diversity of Advanced Sentence Structures (DASS) in writing predicts argumentative writing quality and receptive academic language skills of fifth-to-eighth grade students.** *Assess Writ.* 2022; **53**: 100649.
Publisher Full Text

DeVore S, Kyle K: **Assessing syntactic and lexicogrammatical use in second language Mandarin writing samples.** *J Second Lang Writ.* 2023; **60**: 101014.
Publisher Full Text

Dong J, Zhao Y, Buckingham L: **Thirty years of writing assessment: a bibliometric analysis of research trends and future directions.** *Assess Writ.* 2024; **61**: 100862.
Publisher Full Text

Effatpanah F, Baghaei P, Karimi MN: **A Mixed Rasch Model analysis of multiple profiles in L2 writing.** *Assess Writ.* 2024; **59**: 100803.
Publisher Full Text

Eguchi M, Kyle K: **L2 collocation profiles and their relationship with vocabulary proficiency: a learner corpus approach.** *J Second Lang Writ.* 2023; **60**: 100975.
Publisher Full Text

Emmanouil K, Zoe G, Christos A, *et al.*: **Linguistic factors affecting L1 language evaluation in argumentative essays of students aged 16 to 18 attending secondary education in Greece.** *Assess Writ.* 2024; **60**: 100844.
Publisher Full Text

Ferrari A: **Linguistica del testo. Principi, fenomeni, strutture.** Carocci, 2014.
Reference Source

Fladung IT, Gruhn S, Österbauer V, *et al.*: **Assessing writing in fourth grade: rhetorical specification effects on text quality.** *Assess Writ.* 2023; **57**: 100764.
Publisher Full Text

Flower L, Hayes JR: **A cognitive process theory of writing.** *Coll Compos Commun.* 1981; **32**(4): 365–387.
Publisher Full Text

Frederiksen JR, Collins A: **A systems approach to educational testing.** *Educ Res.* 1989; **18**(9): 27–32.
Publisher Full Text

Ghaffar MA, Khairallah M, Salloum S: **Co-constructed rubrics and assessment for learning: the impact on middle school students' attitudes and writing skills.** *Assess Writ.* 2020; **45**: 100468.
Publisher Full Text

Gong T, Zhang M, Li C: **Association of keyboarding fluency and writing performance in online-delivered assessment.** *Assess Writ.* 2022; **51**: 100575.
Publisher Full Text

Hadidi A: **Comparing Summative and Dynamic Assessments of L2 written argumentative discourse: microgenetic validity evidence.** *Assess Writ.* 2023; **55**: 100691.
Publisher Full Text

Hao Y, Jin Z, Yang Q, *et al.*: **To predict L2 writing quality using lexical richness indices: an investigation of learners of Chinese as a Foreign Language.** *System.* 2023; **118**: 103123.
Publisher Full Text

Hao Y, Wang X, Bin S, *et al.*: **How Syntactic Complexity indices predict Chinese L2 writing quality: an analysis of unified dependency syntactically-annotated corpus.** *Assess Writ.* 2024; **61**: 100847.
Publisher Full Text

He C, Zeng J, Chen J: **Students' motivation for rubric use in the EFL classroom assessment environment.** *Front Psychol.* 2022; **13**: 895952.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Imbler AC, Clark SK, Young TA, *et al.*: **Teaching second-grade students to write science expository text: does a holistic or analytic rubric provide more meaningful results?** *Assess Writ.* 2023; **55**: 100676.
Publisher Full Text

Jacobs HL, Zingraf SA, Wormuth DR, *et al.*: **Testing ESL composition. A practical approach.** Newbury House, 1981.
Reference Source

Jonsson A, Svingby G: **The use of scoring rubrics: reliability, validity and educational consequences.** *Educ Res Rev.* 2007; **2**(2): 130–144.
Publisher Full Text

Kaplan RB: **Cultural thought patterns in inter-cultural education.** *Lang Learn.* 1966; **16**(1–2): 1–20.
Publisher Full Text

Keller SD, Lohmann J, Trüb R, *et al.*: **Language quality, content, structure: what analytic ratings tell us about EFL writing skills at upper secondary school level in Germany and Switzerland.** *J Second Lang Writ.* 2024; **65**: 101129.
Publisher Full Text

Khushik GA: **Is the variation in syntactic complexity features observed in argumentative essays produced by B1 level EFL learners in Finland and Pakistan attributable exclusively to their L1?** *Assess Writ.* 2024; **60**: 100839.
Publisher Full Text

Kim M, Crossley SA: **Modeling second language writing quality: a structural equation investigation of lexical, syntactic, and cohesive features in source-based and independent writing.** *Assess Writ.* 2018; **37**: 39–56.
Publisher Full Text

Kim S, Kessler M: **Examining L2 English university students' uses of lexical bundles and their relationship to writing quality.** *Assess Writ.* 2022; **51**: 100589.
Publisher Full Text

Kim H, Ro E: **Usage-based approaches to assessing syntactic sophistication in second language writing: interaction of genre and proficiency.** *J Second Lang Writ.* 2024; **65**: 101131.
Publisher Full Text

Kim M, Tian Y, Crossley SA: **Exploring the relationships among cognitive and linguistic resources, writing processes, and written products in second language writing.** *J Second Lang Writ.* 2021; **53**: 100824.
Publisher Full Text

Kobayashi H, Rinnert C: **Factors affecting composition evaluation in an EFL context: cultural rhetorical pattern and readers'background.** *Lang Learn.* 1996; **46**(3): 397–433.
Publisher Full Text

Kobrin JL, Deng H, Shaw EJ: **The association between SAT prompt characteristics, response features, and essay scores.** *Assess Writ.* 2011; **16**(3): 154–169.
Publisher Full Text

Kyle K, Crossley SA, Jarvis S: **Assessing the validity of lexical diversity indices using direct judgements.** *Lang Assess Q.* 2021; **18**(2): 154–170.
Publisher Full Text

Leki I: **Good writing: i know it when I see it.** In: *Academic writing in a second language: essays on research and pedagogy.* Ablex Pub. Corp, 1995; 23–46.

Li H, Yu X: **Verb Argument Constructions in argumentative essays by college-level Asian learners of English: exploring the effects of English proficiency, acquisition context, and topic.** *J Second Lang Writ.* 2024; **65**: 101127.
Publisher Full Text

Li J, Huang J: **The impact of essay organization and overall quality on the holistic scoring of EFL writing: perspectives from classroom English teachers and national writing raters.** *Assess Writ.* 2022; **51**: 100604.
Publisher Full Text

Li XM: **"Good Writing" in cross-cultural context.** State University of New York

Press, 1996.
Reference Source

Li Y, Yang R: **Assessing the writing quality of English research articles based on absolute and relative measures of syntactic complexity.** *Assess Writ.* 2023; **55**: 100692.
Publisher Full Text

Li YK, Lin S, Liu Y, *et al.*: **The predictive powers of fine-grained syntactic complexity indices for letter writing proficiency and their relationship to pragmatic appropriateness.** *Assess Writ.* 2023; **56**: 100707.
Publisher Full Text

Liao J: **What skills are being assessed? Evaluating L2 Chinese essays written by hand and on a computer keyboard.** *Assess Writ.* 2023; **57**: 100765.
Publisher Full Text

Lines H: **Chapter 10. "A matter of personal taste": teachers' constructs of writing quality in the secondary school english classroom.** In: C. Bazerman, P. Rogers, A. Stansell, C. Dean, J. S. Early, K. Lunsford, & S. Null (Eds.), *International Advances in Writing Research: Cultures, Places, Measures.* The WAC Clearinghouse; Parlor Press, 2012; 167–187.
Publisher Full Text

Liu Y, Liu J: **Comparing computer-based and paper-based rating modes in an English writing test.** *Assess Writ.* 2023; **57**: 100771.
Publisher Full Text

Lyashevskaya O, Panteleeva I, Vinogradova O: **Automated assessment of learner text complexity.** *Assess Writ.* 2021; **49**: 100529.
Publisher Full Text

Maamuujav U: **Examining lexical features and academic vocabulary use in adolescent L2 students' text-based analytical essays.** *Assess Writ.* 2021; **49**: 100540.
Publisher Full Text

Maamuujav U, Olson CB, Chung H: **Syntactic and lexical features of adolescent L2 students' academic writing.** *J Second Lang Writ.* 2021; **53**: 100822.
Publisher Full Text

MacArthur CA, Jennings A, Philippakos ZA: **Which linguistic features predict quality of argumentative writing for college basic writers, and how do those features change with instruction?** *Read Writ.* 2019; **32**(6): 1553–1574.
Publisher Full Text

McCarthy KS, Roscoe RD, Allen LK, *et al.*: **Automated writing evaluation: does spelling and grammar feedback support high-quality writing and revision?** *Assess Writ.* 2022; **52**: 100608.
Publisher Full Text

Moore N: **What's the point? The role of punctuation in realising information structure in written English.** *Functional Linguist.* 2016; **3**(1): 6.
Publisher Full Text

Mostafa T, Crossley SA: **Verb argument construction complexity indices and L2 writing quality: effects of writing tasks and prompts.** *J Second Lang Writ.* 2020; **49**: 100730.
Publisher Full Text

Mayring P: **Qualitative content analysis: theoretical foundation, basic procedures and software solution.** *Gesis.* 2014.
Reference Source

Naismith B, Attali Y, LaFlair GT: **The impact of task duration on the scoring of independent writing responses of adult L2-English writers.** *Assess Writ.* 2024; **62**: 100895.
Publisher Full Text

Nguyen LQ, Nguyen BTT, Phuong HY: **Exploring the use of model texts as a feedback instrument in expository writing: EFL learners' noticing, incorporations, and text quality.** *Assess Writ.* 2024; **62**: 100890.
Publisher Full Text

Pallotti G: **CAF: defining, refining and differentiating constructs.** *Appl Linguist.* 2009; **30**(4): 590–601.
Publisher Full Text

Peng Y, Sun J, Quan J, *et al.*: **Predicting Chinese EFL learners' human-rated writing quality in argumentative writing through multidimensional computational indices of lexical complexity.** *Assess Writ.* 2023; **56**: 100722.
Publisher Full Text

Peng Y, Zheng Y, Sun J, *et al.*: **Modeling relationships among large-grained, fine-grained absolute syntactic complexity and assessed L2 writing quality: an SEM approach.** *Assess Writ.* 2024; **61**: 100875.
Publisher Full Text

Pollitt A: **Comparative judgement for assessment.** *Int J Technol Des Educ.* 2012; **22**(2): 157–170.
Publisher Full Text

Pun J, Li WK: **A structural equation investigation of linguistic features as indices of writing quality in assessed secondary-level EMI learners' scientific reports.** *Assess Writ.* 2024; **62**: 100897.
Publisher Full Text

Qian L: **Use of lexical features in high-stakes tests: evidence from the perspectives of complexity, accuracy and fluency.** *Assess Writ.* 2023; **57**: 100758.
Publisher Full Text

Riemenschneider A, Weiss Z, Schröter P, *et al.*: **Linguistic complexity in teachers' assessment of German essays in high stakes testing.** *Assess Writ.* 2021; **50**: 100561.
Publisher Full Text

Rock K: **Constructing a data-based analytic rubric for an academic blog post.** *Assess Writ.* 2022; **51**: 100602.
Publisher Full Text

Schmidgall J, Powers DE: ***TOEIC*® Writing test scores as indicators of the functional adequacy of writing in the international workplace: evaluation by linguistic laypersons.** *Assess Writ.* 2020; **46**: 100492.
Publisher Full Text

Schoonen R: **Generalizability of writing scores: an application of structural equation modeling.** *Language Testing.* 2005; **22**(1): 1–30.
Publisher Full Text

Shafiee Rad H, Alipour R: **Unlocking writing success: building assessment literacy for students and teachers through effective interventions.** *Assess Writ.* 2024; **59**: 100804.
Publisher Full Text

Siekmann L, Parr JM, Busse V: **Structure and coherence as challenges in composition: a study of assessing less proficient EFL writers' text quality.** *Assess Writ.* 2022; **54**: 100672.
Publisher Full Text

Tan TX, Fan X, Braunstein LB, *et al.*: **Linguistic, cultural and substantive patterns in L2 writing: a qualitative illustration of MisLevy's sociocognitive perspective on assessment.** *Assess Writ.* 2022; **51**: 100574.
Publisher Full Text

Teng LS: **Explicit strategy-based instruction in L2 writing contexts: a perspective of self-regulated learning and formative assessment.** *Assess Writ.* 2022; **53**: 100645.
Publisher Full Text

Teng LS, Zhang LJ: **Empowering learners in the second/foreign language classroom: can self-regulated learning strategies-based writing instruction make a difference?** *J Second Lang Writ.* 2020; **48**: 100701.
Publisher Full Text

Thwaites P, Kollias C, Paquot M: **Is CJ a valid, reliable form of L2 writing assessment when texts are long, homogeneous in proficiency, and feature heterogeneous prompts?** *Assess Writ.* 2024; **60**: 100843.
Publisher Full Text

Uzun K: **Performance prediction strengths of noun and verb phrases in L2 writing: comparison of density and complexity variables.** *Assess Writ.* 2021; **50**: 100572.
Publisher Full Text

Valtin R, Bird V, Brooks G, *et al.*: **European declaration of the right to literacy.** *European Literacy Policy Network (ELINET).* 2016.
Reference Source

Vandermeulen N, Lindgren E, Waldmann C, *et al.*: **Getting a grip on the writing process: (Effective) approaches to write argumentative and narrative texts in L1 and L2.** *J Second Lang Writ.* 2024; **65**: 101113.
Publisher Full Text

Vasylets O, Marín J: **The effects of working memory and L2 proficiency on L2 writing.** *J Second Lang Writ.* 2021; **52**: 100786.
Publisher Full Text

Wang L, Fan J: **Assessing business english writing: the development and validation of a proficiency scale.** *Assess Writ.* 2020; **46**: 100490.
Publisher Full Text

Wang Y, Xie Q: **Diagnosing EFL undergraduates' discourse competence in academic writing.** *Assess Writ.* 2022; **53**: 100641.
Publisher Full Text

Weigle SC: **Assessing Writing.** (1st ed.) Cambridge University Press, 2002.
Publisher Full Text

Wolf K, Stevens E: **The role of Rubrics in advancing and assessing student learning.** 2007; **7**(1): 3–14.
Reference Source

Woods K, Hashimoto B, Brown EK: **A multi-measure approach for lexical diversity in writing assessments: considerations in measurement and timing.** *Assess Writ.* 2023; **55**: 100688.
Publisher Full Text

Yang Y, Yap NT, Mohamad Ali A: **Predicting EFL expository writing quality with measures of lexical richness.** *Assess Writ.* 2023; **57**: 100762.
Publisher Full Text

Yasuda S: **Does "more complexity" equal "better writing"? Investigating the relationship between form-based complexity and meaning-based complexity in high school EFL learners' argumentative writing.** *Assess Writ.* 2024; **61**: 100867.
Publisher Full Text

Zhang M, Chen W: **Assessing collaborative writing in the digital age: an exploratory study.** *J Second Lang Writ.* 2022; **57**: 100868.
Publisher Full Text

Zhang X, Lu X: **Revisiting the predictive power of traditional vs. fine-grained syntactic complexity indices for L2 writing quality: the case of two genres.** *Assess Writ.* 2022; **51**: 100597.
Publisher Full Text

Zhang Y, Ouyang J: **Linguistic complexity as the predictor of EFL independent and integrated writing quality.** *Assess Writ.* 2023; **56**: 100727.
Publisher Full Text

Open Peer Review

Current Peer Review Status: ? ✓ ✗

Version 1

Reviewer Report 08 January 2026

https://doi.org/10.21956/openreseurope.20902.r65959

© **2026 Balade J.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

✗ **Jennifer Balade**
University of La Laguna, San Cristóbal de La Laguna, Canary Islands, Spain

I really appreciate the opportunity to review the review article entitled *"The measurement of text quality: current methods and open challenges."* Overall, the manuscript addresses an issue of clear relevance for both the educational and scientific communities, namely the inherent complexity of operationalizing and measuring written text quality. The structure of the review generally aligns with the sections proposed in the journal guidelines (Introduction, Literature Analysis, and Conclusion). However, the transition between the theoretical rationale and the literature analysis occasionally appears somewhat blurred. To enhance readability, it may be helpful to more clearly delineate where the introductory framework ends and where the literature analysis begins, perhaps by adopting a structure that minimizes overlap between conceptual definitions and the findings derived from the review itself.
With respect to the theoretical framework, the introduction appropriately emphasizes the importance of writing and its sociocultural variability, yet it could benefit from a deeper linguistic perspective. A substantial body of literature suggests that, although written production involves universal processes, differences in orthographic transparency introduce important distinctions. For instance, assessing text quality in transparent orthographies such as Spanish or Italian entails different challenges and processing demands than in more opaque systems such as English. Incorporating this distinction would contribute to a more robust and contextually grounded theoretical foundation. In addition, the absence of references to previous review articles with similar objectives is noticeable. Situating the present work in relation to existing syntheses would help to better justify its contribution and clarify the specific gap that this review seeks to address.
A further aspect that would benefit from clarification concerns the relationship between the title and the actual scope of the review. While the title suggests a broad overview of text quality measurement methods, the analysis focuses predominantly on studies conducted in English. Although studies in Greek or Chinese are briefly mentioned, the second research question explicitly restricts the analysis to English. It may therefore be advisable to clarify this linguistic delimitation already in the abstract and introduction, in order to avoid potential ambiguity regarding the generalizability of the conclusions. Additionally, to further strengthen the academic tone of the manuscript, some expressions that diverge from standard scholarly style—such as *"as more or less"* (p. 3), *"when looking at the 20 studies,"* or *"in the studies analysed here"* (p. 5)—could be

revised and replaced with more precise formulations. It would also be advisable to carefully check compliance with APA style conventions, for example in expressions such as *"comparative judgments (CJ) (Pollitt, 2012)"* on page 4.

The methodological section arguably presents the most substantial challenges. The decision to restrict the literature search to only two specialized journals, based on Dong et al. (2024), represents a methodological limitation that warrants careful consideration. Given that Dong also identifies eight additional relevant journals, the resulting selection bias may be considerable and could partly explain the limited linguistic diversity observed in the review. A more detailed justification of this decision would strengthen the methodology, and explicitly acknowledging it as an important limitation in the conclusions would enhance transparency. Furthermore, to ensure replicability, it would be important to clearly specify the inclusion and exclusion criteria that led from an initial pool of 134 articles to the final sample of 60 studies.

Regarding the presentation of the findings, the initial subsection describing how text quality is measured includes detailed explanations of rating scales and comparative judgments which, due to their instructional nature, might be more effectively placed in the introduction so as not to disrupt the flow of the literature analysis. In addition, the use of these scales would benefit from greater contextualization. Specifying whether the reviewed studies employ these tools in psychometric research, intervention studies, or preventive frameworks (such as response-to-intervention models) would provide valuable clarity, particularly given that holistic and analytic assessments may pose different challenges across contexts. Similarly, including information on the age ranges of the samples—either in the tables or in the text—would help readers identify which instruments are most appropriate for different developmental stages, from childhood to adulthood. In this regard, greater consistency in the use of acronyms across Tables 1 and 2 (e.g., using TOEFL or DET uniformly) would also improve clarity, as would a clearer explanation of the specific role of Tables 3 and 4 within the narrative.

Finally, although the conclusion formally addresses the research questions, it appears somewhat limited in scope. The overall picture emerging from the review suggests a strong reliance on analytic and holistic scales in L2 contexts, alongside relatively limited attention to intra-rater reliability and to the validation of instruments in languages other than English. Additional detail regarding specific assessment batteries would further strengthen this section, as would a more explicit discussion of the study's limitations. Overall, while the topic addressed is timely and relevant, the manuscript in its current form presents several substantial limitations that limit its contribution to the field. In particular, concerns related to the scope of the literature search and the transparency of key methodological decisions reduce the strength and generalizability of the conclusions. For these reasons, the article would require extensive revision before it could meet the standards expected for indexing as a review article in this journal.

Does the review offer a comprehensive analysis of the research literature?

No

Are all statements factually correct and adequately supported by citations?

Yes

Is the review written in clear language?

Yes

Are the conclusions drawn appropriate in the context of the current literature, books and book chapters?

No

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: early writing development, developmental psychology, specific learning disabilities

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Reviewer Report 31 December 2025

https://doi.org/10.21956/openreseurope.20902.r65962

© 2025 Kübra Er H. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Haticetül Kübra Er
Erzurum Technical University, School of Foreign Languages, Erzurum, Turkey

1. The title is clear and informative, providing a good overview of the research focus. However, it could be made more concise for clarity.
2. Data Analysis: The description of data analysis is clear. It is clearly mentioned whether is this a QUAL or QUAN methodology.
3. The discussion effectively interprets the findings, but it could delve deeper into the implications, especially regarding how these findings could impact English language teaching practices in the specific context.
4. Suggestions for further research or limitation part could be added.

References

1. Kırmızı Ö, Er H: The Champions of If-Constructions in Literary Studies. 2025. 307-326 Publisher Full Text

Does the review offer a comprehensive analysis of the research literature?
Yes

Are all statements factually correct and adequately supported by citations?
Yes

Is the review written in clear language?
Yes

Are the conclusions drawn appropriate in the context of the current literature, books and

book chapters?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: ENGLISH LANGUAGE TEACHING, ENGLISH FOR ACADEMIC AND SPECIFIC PURPOSES, ASSESSING WRITING

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 12 April 2025

https://doi.org/10.21956/openreseurope.20902.r52952

© 2025 Wang H. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Heqiao Wang
Michigan State University, East Lansing, USA

Thank you for the opportunity to review the article titled *"The Measurement of Text Quality: Current Methods and Open Challenges."* Overall, this is a comprehensive and well-organized literature review that effectively summarizes various aspects of writing measurement, including methodologies, rubrics, and evaluation metrics. The article is generally easy to follow, and I appreciate how the authors structured the review around three critical questions in the field. I also appreciate the authors for addressing the complex relationship between text quality and writing proficiency, which remains an important and unresolved issue in writing research.
Below are my suggestions for strengthening the manuscript:

1. After reading the whole article, I feel the claim that text quality does not necessarily equate to writing proficiency is a foundational issue for your paper. I encourage the authors to expand on this point with more support from prior literature. This would help provide a more solid theoretical basis for the analysis. Also, many of the cited sources in the literature review are outdated. Include more recent studies during the last decade would help situate your work within current scholarly conversations and enhance its relevance to contemporary audiences.
2. This analysis is based on only two major journals, resulting in a total of 134 articles. While I understand the rationale for this selection—particularly if you are the sole coder—it does raise concerns about representativeness. I suggest that this limitation be more explicitly acknowledged and discussed in the conclusion, particularly as a direction for future research.
3. After identifying 134 articles, you mention selecting 60 for further analysis. However, the criteria for this selection are unclear to me. Please clarify how these 60 papers were chosen. Include your selection criteria in the article. Additionally, since you include the *Journal of Second Language Writing*, I assume many of the studies focus on L2 learners. Do you also include research on native English-speaking students or general education contexts? Is this

your intention or because of the small scope size? Clarify participant characteristics—such as age, grade level, and language background, because as I read, I feel a confused about your selection standards and participants. A clearly defined selection process is needed.
4. Consider moving the footnote into the main text. This is a very good explanation. The rationale for conducting this literature review is important.
5. In addressing RQ3, the current analysis lists evaluation metrics used in various studies, but to me, this alone does not sufficiently answer how the validity and reliability of writing assessments are ensured. I recommend including IRR statistics/other types of evaluation metrics, if available. Maybe add an additional column to Table 6 related to IRR stats.

Does the review offer a comprehensive analysis of the research literature?
Partly

Are all statements factually correct and adequately supported by citations?
Partly

Is the review written in clear language?
Yes

Are the conclusions drawn appropriate in the context of the current literature, books and book chapters?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: writing measurement and assessment

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.