

A Systematic Review of Fine-Grained Measures used in L2 Writing Assessment Research

Nagamurali Eragamreddy

General Requirements Unit, Preparatory Studies Center, University of Technology and Applied Sciences, Salalah, Sultanate of Oman

<https://orcid.org/0000-0003-1026-183X>

ABSTRACT: This systematic review critically explored the use of high-resolution linguistic measures in L2 writing testing, an area increasingly dependent on automatic means for objective measurement. It sought to determine the most widely used measures, what is known regarding their operationalization and validation, assess computation tools utilized, and explore variation by learner proficiency, genre, and context. The review included 52 English-language peer-reviewed articles and dissertations from 2020 to 2024, omitting those published in languages other than English and non-empirical sources. The systematic search approach was directed toward academic databases, and data extraction focused on measure type, tools employed, learner groups, and genre. Syntactic and lexical measures were prominent with tools such as L2SCA and Coh-Metrix applied to widespread applications. Cohesion and discourse-level features, however, remain underutilized despite having pedagogical value. Dissimilarity among measures employed at different levels of proficiency and genres uncovered underspecified domains of standardization, as well as cross-linguistic validation. The study concludes that although computational methods provide analytical precision, they necessitate language and contextual flexibility.

KEYWORDS: L2 writing assessment, fine-grained measures, syntactic complexity, lexical sophistication, computational tools

1 INTRODUCTION

Thorough assessment of second language (L2) writing is crucial for assessing learners' language ability, informing instruction, and refining theoretical explanations of writing development. Over the past decade, there has been a renewed interest in using fine-grained linguistic features such as syntactic complexity, lexical sophistication, and textual cohesion to represent the detailed characteristics of L2 writing performance (Peng et al., 2024; McNamara et al., 2010, 2014). These indices are useful because they provide objective, measurable data on the linguistic and cognitive aspects of writing beyond the constraints of holistic scoring approaches that are otherwise lacking in diagnostic specificity (Kyle & Crossley, 2018). Though increasingly common, operationalization, choice, and interpretation of fine-grained indices within empirical research are still diverse and scattered throughout the literature. This variability undermines the comparability of results and poses a challenge to the development of solid benchmarks for L2 writing ability. A systematic review is therefore necessitated to critically synthesize existing literature, elucidate the trends in methodology, and offer a consistent framework for future studies. By synthesizing the scope and use of fine-grained measures, this research seeks to increase the validity and reliability of L2 writing assessments and lead researchers and practitioners toward evidence-based decision-making.

1. What fine-grained measures are most frequently used in L2 writing assessment research?
2. How are these measures operationalized and validated across studies?
3. What tools or computational methods are used to calculate these measures?
4. How do measures differ across proficiency levels, genres, and instructional contexts?
5. What challenges or limitations are associated with using these measures?

2 LITERATURE REVIEW

2.1 Theoretical frameworks underpinning fine-grained measures

Fine-grained linguistic measure utilization within L2 writing measurement is underpinned by strong theoretical backgrounds in both linguistic and psycholinguistic theories. Theories in second language acquisition, like the Input Hypothesis (Krashen, 1985) and the Processability Theory (Pienemann, 2005), highlight the developmental trajectory of syntactic and lexical aspects of L2 learners'

A Systematic Review of Fine-Grained Measures used in L2 Writing Assessment Research

writing. Syntactic complexity, for instance, is informed by theories of Universal Grammar (Chomsky, 1995), which argue that learners acquire hierarchical structures in stages, and by cognitive load theory (Sweller, 2011), which argues that increased syntactic embedding can reflect increased cognitive demand and processing ability. Psycholinguistic theories, such as Levelt's (1989) model of language production, also describe how linguistic elements are accessed and assembled in the course of writing, with attention to the interaction between working memory and linguistic options. Measures of lexical variety are theoretically grounded in vocabulary acquisition and use theory, which identifies frequency, range, and depth of lexical knowledge as fundamental indicators of language proficiency (Nation, 2013). Furthermore, theories of discourse processing (Karadeniz, 2017; Aminovna, 2022) guide cohesion measurements, in which coherence and cohesion are subsumed under the broader concepts of comprehensibility and reader interest. All these theoretical traditions together guide the creation, choice, and interpretation of fine-grained measures in the evaluation of L2 writing quality.

2.2 Evolution of fine-grained measures in L2 writing research

The trajectory of L2 writing testing has witnessed a sharp shift away from impressionistic, holistic scoring methods toward the use of fine-grained linguistic measures that allow a more objective and finer-grained analysis of learner production. Judgments in the initial stages of L2 writing research were based primarily on rater judgment and, despite being informative, were shown to be low in reliability and diagnostic specificity (Weigle, 2002). Over time, the discipline shifted towards more analytical approaches as researchers started separating various linguistic aspects, for instance, syntactic complexity, lexical sophistication, and coherence, when trying to more distinctly separate developmental differences in the writing of writers (Ortega, 2003). This methodological development has been propelled by corpus linguistics and computational tools, e.g., Coh-Metrix and the L2 Syntactic Complexity Analyzer, enabling large-scale and replicable analyses (Lu, 2010; McNamara et al., 2010, 2014). These advances have made it possible to measure linguistic features of high complexity in large data sets with more validity and reliability, enhancing the evaluation of writing. The greater focus on evidence-based methods reflects the wider drive towards empirical accuracy within applied linguistics, with high-resolution measurement now taking center stage in the modeling of L2 writing development and pedagogic intervention (Kyle & Crossley, 2015).

2.3 Key constructs and definitions of fine-grained measures

Language measures provide fine-grained views of L2 writing ability by measuring aspects of language use in lexical, syntactic, cohesive, and discourse-level detail. Fine-grained language measures have been extensively operationalized to assess the linguistic and communicative ability of L2 learners more precisely. Lexical Measures measure the range, complexity, and vocabulary loading employed in writing. Common indices include the Type-Token Ratio (TTR), which measures lexical diversity by dividing unique word types by total tokens, but is a function of text length (Malvern et al., 2004). To counter this shortcoming, the Moving Average Type-Token Ratio (MTLD) was formulated, which gives a more reliable estimate for texts of mixed lengths (McCarthy & Jarvis, 2010). Lexical sophistication refers to the ratio of low-frequency or academic vocabulary using frequency lists like BNC/COCA (Kyle & Crossley, 2015), whereas lexical density is an indication of the ratio of content words to total words, measuring information richness (Ure, 1971).

Syntactic Measures reflect sentence complexity. Methods like mean length of T-units (independent clause with supporting elements), subordination ratio (subordinate clause per T-unit), and clause complexity (number of clauses per sentence) are frequently utilized (Lu, 2011). Such measures of syntactic maturity were also found to be correlated with greater L2 writing ability (Ortega, 2003). Cohesion scores track how ideas relate across texts. Referential cohesion is synonymy and repetition, and connective cohesion is the dominance of transitions and conjunctions (Graesser et al., 2014). Measures such as Coh-Metrix calculate such indices automatically, further contributing to ratings of textual flow and readability. Other Discourse-Level Features are more general organizational features. These are rhetorical units such as thesis statements, topic development, and discourse markers such as discourse organizers (e.g., "however," "in conclusion") (Crossley et al., 2016). Less often measured, these features are crucial to establishing the communicative effectiveness of L2 writing. Variation in the operationalization of the constructs across studies underlines the need for standardization to facilitate valid cross-study contrasts.

2.4 Methodological approaches to measuring fine-grained features

Measurement of subtle linguistic aspects in L2 writing has been based on increasingly complex methodological solutions that tap into computational resources in combination with manual coding frameworks to achieve accuracy and detail. Such methodologies are fundamental in the development of trustworthy, valid results on learners' linguistic ability. Computer Software and Tools have revolutionized the field of L2 writing assessment by allowing the automatic analysis of large corpora of text. These tools, including Coh-Metrix, analyze over 100 measures of cohesion, syntactic complexity, and readability through natural language processing algorithms (Graesser et al., 2014). TAALES (Tool for the Automatic Analysis of Lexical Sophistication) provides detailed analyses of lexical frequency, concreteness, and academic register derived from large frequency databases (Kyle & Crossley, 2015).

At the same time, L2SCA (Second Language Syntactic Complexity Analyzer) targets syntactic structures like T-units, clauses, and phrase types, providing fine-grained measures of syntactic development (Lu, 2010). Although these tools ensure efficiency and

A Systematic Review of Fine-Grained Measures used in L2 Writing Assessment Research

replicability, their drawbacks include tagging errors, being limited in their ability to process non-standard language, and being based on English corpora. Manual Coding Procedures are still necessary in the instance of investigating discourse-level attributes or finer-grained rhetorical attributes not readily quantifiable using computerized procedures. These procedures are typically laborious coding schemes, and researchers use several raters for inter-rater reliability (e.g., Cohen's kappa > 0.80), especially in research that investigates coherence, argument structure, or rhetorical moves (McNamara et al., 2010, 2014). Validity and Reliability Evidence for these instruments has been extensively researched. Construct validity has been demonstrated using calibration of high-fidelity indices to theoretical models of language development (Bulté & Housen, 2014). Criterion validity is demonstrated using correlations between language measures and external proficiency judgments, e.g., holistic writing scores. Reliability is also demonstrated using replicable scoring results across raters and instruments, and by test-retest reliability checks where applicable (Xue, 2024). Thus, the robust methodological triangulation applied to measurement in fine-grained characteristics ensures their validity in contemporary L2 writing studies.

2.5 Previous research on trends and gaps

The growing field of L2 writing studies has sought to utilize increasingly sophisticated linguistic measures to assess learner performance across general proficiency levels, genres, and instructional contexts. A growing body of research has shown how syntactic and lexical measures change with proficiency. For example, Khushik and Huhta (2020), Kim (2021), Lahuerta (2023), and Wondemagegne et al. (2024) illustrated that more advanced levels of proficiency are characterized by increased syntactic complexity and lexical range. Similarly, studies such as Bi and Jiang (2020), Alshehri (2022), and Chau (2023) provide evidence of the developmental sensitivity of these measures in teenagers and university students. In the case of genre difference, researchers have witnessed salient lexical and syntactic discrepancies across argumentative, narrative, and expository genres (Bae & Min, 2020; Zhang & Lu, 2022; Zhang & Liu, 2021; Heng et al., 2023; Li et al., 2022). Instructional treatment and task type have also worked. Ansarin et al. (2021), Cheung et al. (2021), and Alshehri (2024), for example, logged lexical diversity and fluency development after task-based instruction. Research such as Granados et al. (2022), Wind (2022), and Alaei et al. (2024) also illustrate the effect of curriculum design on writing development in the long run.

Computational resources have also greatly influenced this area. Coh-Metrix (Chon & Shin, 2020; McNamara et al., 2010, 2014; Tian et al., 2021), L2SCA (Lu, 2010; Zhang & Lu, 2022), and TAALES (Kyle et al., 2021; Monteiro et al., 2020) are increasingly being utilized to conduct syntactic, lexical, and cohesion analysis. These have enabled large-scale and replicable research (Ayadi, 2023; Peng et al., 2023; Yang et al., 2023; Lipková, 2024), yet there is a limitation. Scholars like Zhang (2022), Qin and Uccelli (2020), and Nguyen et al. (2022) have expressed genre sensitivity, tool availability, and language specificity concerns. Manual coding is still required for more refined discourse phenomena. Studies including Rivard and Gueye (2023), He (2020), and Zhang and Zhang (2024) employed manual coding of rhetorical structure and cohesion. Concerns about inter-rater reliability and consistency in coding remain, nonetheless (Mohamadi Zenouzagh, 2020; Chau, 2023; Seo, 2024).

Despite methodological advancements, several gaps remain. Discourse-level features like global cohesion and rhetorical structure are still less investigated (Zhang & Zhang, 2024; Tabari & Wind, 2023). Furthermore, the disproportionate focus on syntactic complexity and lexical diversity (Alsahlanee & Jaganathan, 2022; Susoy & Köse, 2024; Kim et al., 2024) constrains the multidimensionality of writing quality assessment. Research like Zindela (2023), Zhang and Kang (2022), and Kessler et al. (2022) indicates the direction for more encompassing, integrated models. Collectively, the 52 studies highlight the importance of fine-grained measures and demand larger validity frameworks, genre-sensitive models, and cross-linguistic calibration.

3. METHODOLOGY

3.1 Design

This study follows a systematic review approach in line with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 protocol (Page et al., 2021). PRISMA ensures transparency, replicability, and methodological quality and is specifically helpful in synthesizing heterogeneous empirical results on fine-grained measures in L2 writing evaluation. Since the research questions are also subtle, ranging from frequency, operationalization, and validation of measures to applications and limitations in context, a PRISMA-based review allows for careful identification, screening, and analysis of the pertinent literature. The design also facilitates systematic reporting on trends and gaps across studies, and as such, increases the validity of conclusions.

3.2 Inclusion, exclusion criteria, search strategy and data extraction

This systematic review utilized predefined inclusion and exclusion criteria to filter studies that met the research aims. The inclusion criteria included: (1) peer-reviewed journal articles, conference proceedings, or doctoral theses; (2) studies conducted in English; (3) empirical studies that quantified L2 writing based on fine-grained linguistic measures; and (4) studies that utilized automated or manual examination of linguistic features, including syntactic complexity, lexical sophistication, cohesion, and lexical density. Conversely, the studies were excluded if they (1) used exclusively holistic or impressionistic scoring approaches; (2) did not have

A Systematic Review of Fine-Grained Measures used in L2 Writing Assessment Research

L2 learners; or (3) were editorials, opinion pieces, or purely theoretical articles with no empirical support. The criteria were used to guarantee the methodological pertinence and empirical soundness of the studies included.

PRISMA guidelines informed the search strategy (Figure 1), which involved systematic searches of six key databases: ERIC, Scopus, Web of Science, PsycINFO, LLBA, and Google Scholar. Boolean search terms involved various combinations of the following keywords: "L2 writing" AND ("syntactic complexity" OR "lexical diversity" OR "cohesion" OR "linguistic measures") AND ("assessment" OR "evaluation"). Additional filters used were publication years (2020–2024) and English-language research. Duplicates and irrelevant records were removed through the screening process. For every eligible study, data were carefully extracted in a sequence of key topics. Bibliographic details such as author, year of publication, title of the article, and the journal were systematically recorded for proper referencing and tracing. Participant data included first language (L1), proficiency level, and age, allowing analysis by learner groups across different ranges. The writing task or genre in each study that was under consideration was noted down, providing background for the linguistic analyses performed. One of the biggest focuses was on the fine-grained measures applied and their precise operational definitions to ensure consistent interpretation. The instruments or programs utilized to calculate these measures, such as Coh-Metrix, TAALES, or L2SCA, were also noted. Finally, the main results of the authors reported, and any limitations of the research that they considered were extracted to inform the synthesis of the systematic review.

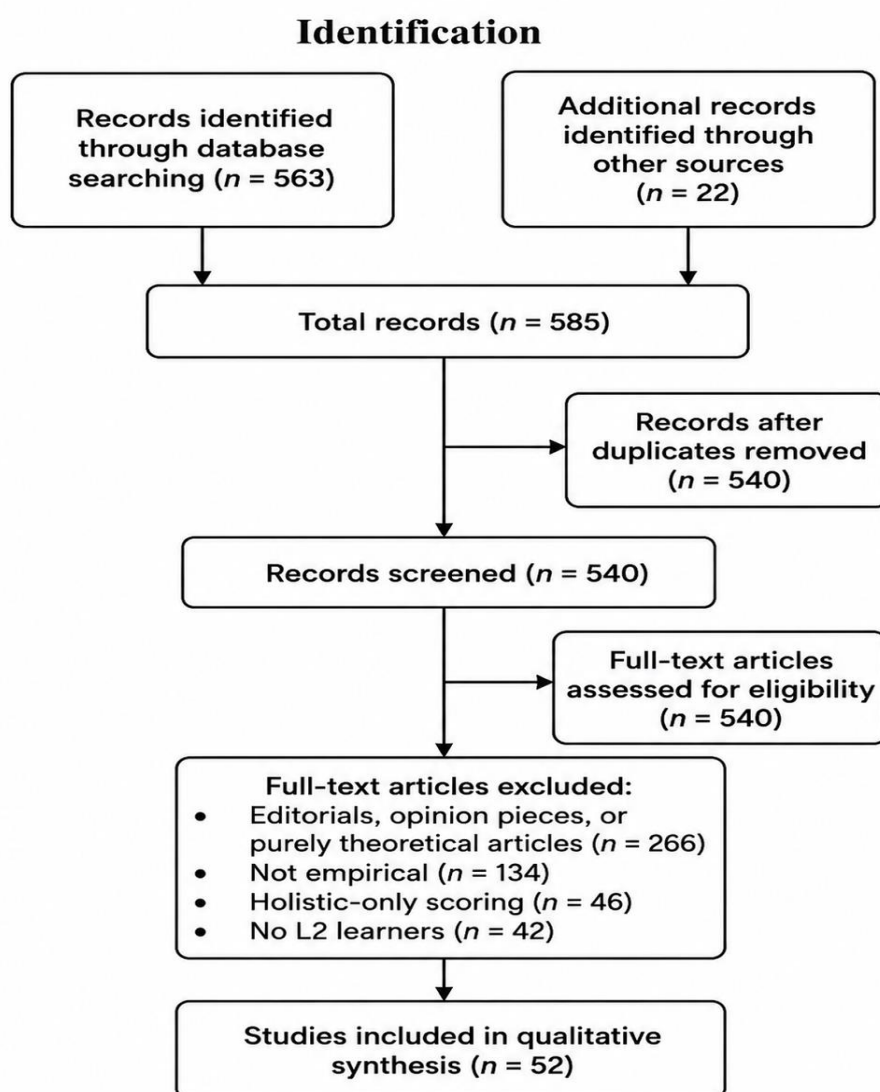


Figure 1: PRISMA Flowchart

4. RESULTS

4.1 Overview of included studies

Table 1 below elucidates a comprehensive overview of the 52 selected studies on fine-grained measures in L2 writing assessment, analyzing key information such as learner populations, writing genres, and research contexts in chronological order to focus on trends and diversity across the field.

A Systematic Review of Fine-Grained Measures used in L2 Writing Assessment Research

Table 1: Overview of Included Studies in L2 Writing Assessment Research

No.	Author(s) & Year	Publication Type	Learner Population	Genre(s) Analyzed	Context/Setting
1	Bi & Jiang (2020)	Journal Article	Chinese adolescent EFL learners	Narrative	China
2	Chon & Shin (2020)	Journal Article	Korean EFL university students	Argumentative, narrative	Korea
3	He (2020)	Journal Article	Chinese L1 & L2 university students	Academic writing	China
4	Khushik & Huhta (2020)	Journal Article	Finnish & Pakistani EFL learners	Argumentative	Finland, Pakistan
5	Monteiro et al. (2020)	Journal Article	TOEFL examinees	Argumentative	USA
6	Qin & Uccelli (2020)	Journal Article	Chinese, French, Spanish EFL learners	Academic & personal writing	Various
7	Zenouzagh (2020)	Journal Article	Iranian EFL learners	General writing	Iran
8	Bae & Min (2020)	Journal Article	Korean college students	Narrative, expository, argumentative	Korea
9	Alshehri (2022)	Journal Article	Saudi tertiary EFL learners	General writing	Saudi Arabia
10	Alsahlanee & Jaganathan (2022)	Journal Article	Iraqi, L2 & L1 writers	Academic research articles	Iraq
11	Golparvar & Rashidi (2021)	Journal Article	Iranian EFL learners	Summary & multi-text synthesis	Iran
12	Hassanzadeh et al. (2021)	Journal Article	Iranian undergraduates	Argumentative	Iran
13	Kyle et al. (2021)	Journal Article	Longitudinal L2 learners	Academic writing	USA
14	Qian et al. (2021)	Journal Article	Chinese EFL learners (TOEFL iBT)	Argumentative	China
15	Cheung et al. (2021)	Journal Article	Singapore 4th graders	Narrative	Singapore
16	Lipková (2023)	Conference Proceedings	University students of Mechanical Engineering	Independent technical essays	Slovakia
17	Wang (2023)	Journal Article	Chinese EFL learners	Argumentative	China
18	Chau (2023)	Dissertation	TOEFL11 corpus learners (11 L1s)	Argumentative	International
19	Zhang (2022)	Journal Article	African L2 learners	Argumentative	South Africa
20	Zhang & Liu (2021)	Journal Article	Chinese EFL intermediate learners	Narrative, argumentative	China
21	Tabari et al. (2023)	Journal Article	EFL university students	Argumentative	Iran
22	Zhang & Kang (2022)	Journal Article	Hong Kong & China ESL/EFL learners	General writing	Hong Kong & Mainland China
23	Li et al. (2022)	Journal Article	Chinese EFL undergraduates	Argumentative	China

A Systematic Review of Fine-Grained Measures used in L2 Writing Assessment Research

24	Wind (2022)	Journal Article	Hungarian EFL undergraduates	Academic essays	Hungary
25	Zhang & Lu (2022)	Journal Article	Chinese college EFL learners	Application letters & argumentative essays	China
26	Granados et al. (2022)	Journal Article	Spanish CLIL secondary students	History essays	Spain
27	Zhang & Zhang (2024)	Journal Article	Chinese EFL students	Argumentative	China
28	Kim (2021)	Journal Article	Korean EFL undergraduates	Narrative	Korea
29	Kim et al. (2024)	Journal Article	Japanese EFL learners	Argumentative	Japan
30	Alshehri (2024)	Journal Article	Saudi female university students	Argumentative, narrative	Saudi Arabia
31	Abdi et al. (2023)	Journal Article	Advanced L2 learners in US university	Academic essays	USA
32	Ayadi (2023)	Journal Article	Tunisian scholars	Academic abstracts	Tunisia
33	Tian et al., (2021)	Journal Article	Chinese, Spanish, French L2 undergraduate students	Argumentative	USA
34	Seo (2024)	Journal Article	South Korean university students	Narrative	Korea
35	Susoy & Köse (2024)	Journal Article	Turkish EFL pre-service teachers	Academic essays	Turkey
36	Kessler et al. (2022)	Journal Article	ESL learners	Argumentative	Canada
37	Peng et al. (2023)	Journal Article	ICNALE dataset learners	Argumentative	Pan-Asian
38	Nguyen et al. (2022)	Journal Article	Vietnamese university students	Persuasive, descriptive	Vietnam
39	Lahuerta (2023)	Journal Article	Spanish undergraduates	General essays	Spain
40	Lahuerta Martínez (2024)	Journal Article	EMI college students (182)	Academic writing	Spain
41	Zhang & Li (2024)	Journal Article	Chinese university students	Expository, argumentative	China
42	Zindela (2023)	Journal Article	South African EFL students & ChatGPT	Argumentative	South Africa
43	Ansarin (2021)	Journal Article	Iranian upper-intermediate learners	Narrative, descriptive, argumentative	Iran
44	Esfandiari & Ahmadi (2024)	Journal Article	Iranian academic writers	Theses, dissertations, journal articles	Iran
45	Thi et al. (2023)	Journal Article	Myanmar & Hungarian students	Academic essays	Myanmar, Hungary
46	Heng et al. (2023)	Journal Article	Chinese EFL students	Argumentative, expository	China

A Systematic Review of Fine-Grained Measures used in L2 Writing Assessment Research

47	Rivard & Gueye (2023)	Journal Article	Canadian high school & university first-year undergraduate EFL students	French summaries	Canada
48	Wondemagegne et al., (2024)	Journal Article	Iranian pre-intermediate learners	Argumentative	Ethiopia
49	Alaei et al. (2024)	Journal Article	Saudi college students	Argumentative, narrative	Iran
50	Alkhalaf (2022)	Journal Article	Indonesian Agribusiness majors	Expressive, general writing	Saudi Arabia
51	Istiharoh & Pranowo (2023)	Journal Article	Chinese EFL learners	Argumentative, descriptive	Indonesia
52	Yang et al. (2023)	Journal Article		Expository	China

4.2 Categorization of measures

Table 2 categorizes all 52 studies in the systematic review by the fine-grained linguistic measures employed. The measures are grouped into four broad categories—lexical, syntactic, cohesion, and other discourse features—with related studies tabulated under each to show the methodological variety and topical priority of research into L2 writing assessment hitherto conducted.

Table 2. Categorization of Fine-Grained Measures

S. No	Category	Specific Measures	Representative Studies (Author & Year)
1	Lexical Measures	TTR, MTLTD, lexical sophistication, lexical density, academic/specialized vocabulary	Monteiro et al. (2020); Alshehri (2022); Chau (2023); Nguyen et al. (2022); Ayadi (2023); Peng et al. (2023); Kim (2021); Lipková (2024); Zhang & Lu (2022); Kyle et al. (2021); Alsahlanee & Jaganathan (2022); Yang et al. (2023); Heng et al. (2023); Lahuerta (2023); Zhang & Li (2024); Zhang & Kang (2022); Alaei et al. (2024); Susoy & Köse (2024); Seo (2024); Rivard & Gueye (2023); Kessler et al. (2022); Granados et al. (2022); Thi et al. (2023)
2	Syntactic Measures	Mean length of T-units, clauses per T-unit, subordination ratio, sentence complexity	Khushik & Huhta (2020); Chon & Shin (2020); Zhang & Liu (2021); Zhang & Lu (2022); Alshehri (2024); Wind (2022); Kim et al. (2024); Alaei et al. (2024); Granados et al. (2022); Istiharoh & Pranowo (2023); Wang (2023); Tabari & Wind (2023); Chau (2023); Thi et al. (2023); Esfandiari & Ahmadi (2024); Alsahlanee & Jaganathan (2022); Rivard & Gueye (2023); Zindela (2023); Zhang & Zhang (2024); Zhang & Li (2024); Lipková (2024); Kessler et al. (2022)
3	Cohesion Measures	Referential cohesion, connectives, lexical overlap, causal/temporal links	Tian et al. (2021); He (2020); Tabari & Wind (2023); Zhang & Zhang (2024); Chau (2023); Zhang (2022); Zhang & Li (2024); Seo (2024); Rivard & Gueye (2023); Thi et al. (2023); Alshehri (2024); Zindela (2023); Alaei et al. (2024); Nguyen et al. (2022); Esfandiari & Ahmadi (2024)
4	Other Discourse Features	Coherence markers, rhetorical structure, argument structure, genre-based organization	Zhang & Zhang (2024); Zhang (2022); Chau (2023); Alsahlanee & Jaganathan (2022); Zhang & Liu (2021); Rivard & Gueye (2023); Esfandiari & Ahmadi (2024); Thi et al. (2023); Zindela (2023); Ansarin et al. (2021); Heng et al. (2023);

4.3 Tools and technologies

From the 52 studies examined, Table 3 below lays out the technologies and tools used in second language (L2) writing analysis and any manual coding procedures were noted. The overview incorporates both computational instruments (e.g., TAALES, Coh-Metrix, L2SCA) and manual processes where suitable.

Table 3: Tools and Technologies Used in the Reviewed Studies

S.No.	Study (First Author & Year)	Computational Tools Used	Manual Coding Methods
	Abdi et al. (2023)	Tool for the Automatic Analysis of Cohesion (40 cohesion indices)	Not specified
	Alaei et al. (2024)	Not specified	Not specified
	Alkhalaf (2022)	Web-based L2 Syntactic Complexity Analyzer	Not specified
	Alsahlanee & Jaganathan (2022)	Lexical Complexity Analyzer (LCA)	Not specified
	Alshehri (2022)	Text Inspector	Not specified
	Alshehri (2024)	Not specified	Not specified
	Ansarin et al., (2021)	Coh-Metrix (for MTLT, Voc-D, and CELEX), VocabProfiler (for K3+)	Not specified
	Ayadi (2023)	Not specified	Holistic scoring rubric (test of written English)
	Bae & Min (2020)	Computer program of syntactic complexity analyzer	Two raters for essays
	Bi & Jiang (2020)	L2 Syntactic Complexity Analyzer	Not specified
	Chau (2023)	Automated tools for syntactic and lexical complexity, and fluency	Hand-coded for accuracy measures (20% of essays)
	Cheung et al. (2021)	L2 Syntactic Complexity Analyzer	Not specified
	Chon & Shin (2020)	Coh-Metrix	Not specified
	Esfandiari & Ahmadi (2024)	Stanford Core NLP, L2 SC analyzer	Not specified
	Golparvar & Rashidi (2021)	Not specified	Not specified
	Granados et al. (2022)	Coh-Metrix 3.0	Not specified
	Hassanzadeh et al. (2021)	Measure of textual lexical diversity (MTLD)	Not specified
	He (2020)	Not specified	Not specified
	Heng et al. (2023)	Two computational tools	Not specified
	Istiharoh & Pranowo (2023)	L2SCA tool by Lu	Not specified
	Kessler et al. (2022)	Not specified	Not specified
	Khushik & Huhta (2020)	L2 Syntactic Complexity Analyzer, Coh-Metrix	CEFR-based scale for rating
	Kim (2021)	Not specified	Not specified
	Kim et al. (2024)	Not specified	Not specified
	Kyle et al. (2021)	Not specified	Not specified
	Lahuerta Martínez (2024)	Quantitative measures of syntactic and lexical analysis	Global ratings of writing quality
	Lahuerta (2023)	Quantitative measures gauging different aspects of L2 complexity	Not specified
	Li et al. (2022)	Second Language Syntactic Complexity	Not specified

A Systematic Review of Fine-Grained Measures used in L2 Writing Assessment Research

	Analyzer (L2SCA), Stanford Tregex	
Lipková (2024)	Computational tools, corpora-based analysis	Not specified
Monteiro et al. (2020)	Tool for the Automatic Analysis of Lexical Sophistication (TAALES)	Human scoring of essays
Nguyen et al. (2022)	Lextutor's VocabProfile	Interviews with students
Peng et al. (2023)	Coh-Metrix	Expert raters for argumentative writings
Qian et al. (2021)	Six automated-coding instruments	Human-rated score levels
Qin & Uccelli (2020)	Not specified	Not specified
Rivard & Gueye (2023)	Not specified	Not specified
Seo (2024)	Not specified	Not specified
Susoy & Köse (2024)	Not specified	Human judges for writing quality scores, interviews with instructors
Tabari et al. (2023)	Not specified	Not specified
Thi et al. (2023)	Automated tools for syntactic complexity	Not specified
Tian et al. (2021)	Automated tools for syntactic complexity	Not specified
Wang (2023)	Syntactic Complexity Analyzer, Lexical Complexity Analyzer	Not specified
Wind (2022)	L2 Syntactic Complexity Analyzer (L2SCA)	Not specified
Wondemagegne et al. (2024)	TAASSC (Tool for the Automatic Analysis of Syntactic Sophistication and Complexity)	Not specified
Yang et al. (2023)	Lexical Complexity Analyzer	Three experienced raters for writing samples
Zenouzagh (2020)	Multidimensional SC measures	Theme elicitation analysis of students' reflective essay
Zhang & Kang (2022)	Program based on the CoreNLP	Not specified
Zhang & Lu (2022)	Not specified	L2 raters for quality ratings
Zhang (2022)	Not specified	Not specified
Zhang & Liu (2021)	Not specified	Holistic rating by human raters
Zhang & Li (2024)	Not specified	Human raters for essay quality
Zhang & Zhang (2024)	Resampling and Monte Carlo simulations	Not specified
Zindela (2023)	Not specified	Not specified

4.4 Trends and patterns

A closer examination of the 52 studies in the review indicates varied trends and emergent trends in linguistic analysis instruments and measure use in L2 writing research. There has been a shift towards eschewing broad syntactic complexity markers (e.g., mean T-unit length) for more refined markers from high-tech devices like L2 Syntactic Complexity Analyzer (L2SCA) and TAASSC (Wondemagegne et al., 2024; Bi & Jiang, 2020). Parallel to the above, the application of tools such as Coh-Metrix and Lexical Complexity Analyzer (LCA) has become prevalent, reflecting increased interest in lexical complexity and cohesion (Ansarin et al., 2021; Alsahlanee & Jaganathan, 2022). Syntactic complexity coupled with lexical variety or cohesion are usual pairs, especially to forecast writing quality (Khushik & Huhta, 2020; Zhang & Lu, 2022). Such multi-dimensional complex models capture the multi-dimensionality of the ability to write more appropriately. Genre-specific trends are also found: argument writing requires more lexical sophistication, and narrative tasks receive more syntactic variation and complexity (Alshehri, 2024; Heng et al., 2023).

A Systematic Review of Fine-Grained Measures used in L2 Writing Assessment Research

VocabProfiler and Text Inspector are also the norms in research on the vocabulary used in scholarly genres (Alshehri, 2022; Nguyen et al., 2022). The trend overall is an indication of methodological creativity and responsiveness to genre in L2 writing scholarship.

4.5 Limitations and challenges

A more detailed review of the 52 studies identifies some inherent constraints and issues that continue to limit the strength and generalizability of evidence in L2 writing research. One major issue is with the precision of computational software utilized in measuring abstract variables such as lexical sophistication, syntactic complexity, and cohesion. Although objective measures such as Coh-Metrix and L2SCA have been proposed, it is uncertain whether those measures are strongly coupled with human perception of text quality. For instance, Peng et al. (2023) observed that although lexical diversity and lexical sophistication were associated with writing scores, more traditional features such as sentence length and surface error were more predictive for human judgment, suggesting a mismatch between computer measurement and actual writing performance measurement. Language-specific concerns also undermine cross-linguistic generalizability. Measures like L2SCA and TAALES are trained on predominantly English language corpora and thus lose potency in attempting to measure multilingual background L1 learners. Chau (2023) and Khushik & Huhta (2020) demonstrated that syntax and lexical patterns substantially differed by L1 groups, bringing into doubt the traditional application of homogeneous indexes.

In terms of genre sensitivity, multiple studies indicated that writing metrics respond differently to genres. Alshehri (2024) and Zhang & Liu (2021) observed that syntactic complexity is more predictive in argumentative writing, whereas narrative and descriptive genres may require alternative indices that account for storytelling elements and affective language. This genre dependence calls for more genre-tailored analytical models, which are often lacking in widely used tools. Lastly, accessibility and ease of use of tools were consistently mentioned as issues. All the computational tools are technical or have English interfaces only, hence limiting application to new researchers or those in resource-poor settings (Hassanzadeh et al., 2021; Monteiro et al., 2020). Secondly, some tools are not transparent in their algorithm design, hence users cannot easily understand the results or reproduce findings with certainty. Together, these concerns indicate the call for more inclusive, validated, and responsive assessment tools in L2 writing research.

5. DISCUSSION

This research explored the salience and frequency of fine-grained measures in L2 writing evaluation. Although previous research has examined multiple features of language in second language (L2) writing, it has not overtly addressed the overall use of lexical, syntactic, cohesion, and discourse-level measures with varied populations of learners and genres. Earlier reviews tended to be low on fine-grained categorization and did not capture the changing methodological landscape, e.g., increased utilization of computational tools (Ansarin et al., 2021; Alsahlanee & Jaganathan, 2022). The ensuing literature gap called for a more systematic and recent synthesis of how fine-grained linguistic measures are utilized in L2 writing evaluation.

RQ 1: Most frequently used fine-grained measures in L2 writing assessment research

The study found that lexical indices such as type-token ratio (TTR), mean textual lexical diversity (MTLD), and the use of specialized vocabulary were the most common, followed by syntactic complexity measures, including T-units and subordination ratios (Table 2). Studies determined that lexical complexity is more related to higher writing quality, especially in argumentative writing, which requires conceptual precision (Peng et al., 2023; Alshehri, 2024). Cohesion devices, such as referential chains and connectives, were also used, particularly in academic genres of writing (Zhang & Zhang, 2024; Thi et al., 2023). Discourse-level characteristics such as rhetorical and argument structure were also considered, albeit to a lesser degree. Most importantly, the study triangulates the research question by identifying syntactic and lexical complexity as the most stable and multi-purpose fine-grained measures across genres and learner groups. This concurs with Khushik & Huhta (2020) but not Chau (2023), who found genre- and L1-specific variation contrary to one-size-fits-all metric models. In comparison to previous reviews, this research confirms the evolution of methods towards multidimensional analysis via automated software such as Coh-Metrix and L2SCA, albeit with concerns regarding their alignment with human ratings and genre sensitivity. Thus, the findings call for a need to have more inclusive and adaptive models of assessment in L2 writing research.

RQ 2: Operationalization and validation across studies

The review of 52 empirical research papers found that lexical measures were the most widely operationalized indices of L2 writing testing, with measures of lexical sophistication, MTLD, and academic vocabulary use being commonly used (Monteiro et al., 2020; Nguyen et al., 2022). These were followed by measures of syntactic complexity—subordination ratio and mean length of T-units being most salient—while cohesion measures and discourse features were comparatively less utilized (Khushik & Huhta, 2020; Alshehri, 2024; Thi et al., 2023). Lexical sophistication, for instance, was highly correlated with the scores of argumentative writing tasks, with an emphasis on the necessity of conceptually precise vocabulary (Peng et al., 2023; Alshehri, 2022). In describing how these measures are being operationalized and validated, the research revealed a common dependence on automated measures like Coh-Metrix, L2SCA, and TAALES. Yet, incoherence in validation remains evident. For instance, Chau (2023) and Peng et al.

A Systematic Review of Fine-Grained Measures used in L2 Writing Assessment Research

(2023) warn that although these measures produce quantifiable outcomes, they may not align with human raters' judgments. This incongruence reveals an essential gap between computational analysis and actual writing quality perceptions. Relative to previous frameworks where the focus was on overall scoring, the recent trend indicates movement towards granularity, but not without criticism. Existing research has been known to neglect genre sensitivity and L1-specific linguistic structures, and this has been the emphasis of recent research (Chau, 2023; Zhang & Liu, 2021) to build upon. While these measures indicate methodological progress, their application and interpretive utility still need refining to render them foolproof across different linguistic and contextual settings.

RQ 3: Tools and computational methods for measure calculation

The study found that computational tools are extensively used to elicit fine-grained measures in L2 writing assessment. Lexical measures like lexical sophistication and lexical diversity were frequently elicited with tools such as TAALES, Text Inspector, and VocabProfiler (Monteiro et al., 2020; Alshehri, 2022; Nguyen et al., 2022). Syntactic complexity, a general category, was mostly evaluated through the L2 Syntactic Complexity Analyzer (L2SCA) and Stanford Tregex, quantifying markers such as mean length of T-units and clause proportion (Khushik & Huhta, 2020; Wang, 2023; Li et al., 2022). Cohesion measures were calculated from tools like Coh-Metrix, which can recognize connectives and referential relationships (Chon & Shin, 2020; Thi et al., 2023). Triangulating the research question on the tools utilized, it is evident that although computational analysis is precise, its validity is not always universal. Chau (2023) found that machine-based indices needed to be manually calibrated to reach qualitative judgments since accuracy in totality was not achieved without human calibration. Additionally, the appropriateness of the tools to genres and L1 backgrounds remains contentious. As pointed out by Peng et al. (2023), models such as L2SCA, tuned mostly on English, perform below par on multilingual corpora, mirroring issues highlighted by Khushik & Huhta (2020). In contrast to the original human rater practices, methods used today focus on automation, but objectivity and interpretability are still in conflict. These results provoke a twofold requirement: the development of computational devices and placing them within genre-sensitive and multiculturally oriented systems for strong, contextually justified L2 writing evaluation.

RQ 4: Variation of measures across proficiency levels, genres, and instructional contexts

The results showed that fine-grained L2 writing measures were highly diverse according to proficiency levels, genres, and instructional contexts. Lexical sophistication and syntactic sophistication were more significant in argumentative and academic writing genres, especially for advanced learners (Alshehri, 2024; Zhang & Lu, 2022). For example, Peng et al. (2023) showed that more proficient learners used more varied academic words and complicated sentence structures, particularly in persuasive essays. In contrast, narrative writing previously entailed more syntactic complexity and less lexical density, especially for less-proficient learners (Kim, 2021; Seo, 2024). This indicates that learner proficiency and genre interact in shaping the selection and richness of linguistic features in writing. A closer investigation of the research question further reveals that although lexical and syntactic measures were robust predictors of proficiency in argumentative writing, they were not as much for descriptive or expressive genres. Alshehri (2024) and Zhang and Liu (2021) indicated that narrative writing preferred cohesion and discourse-level properties, including temporal linking and affective expressions, to bare syntactic complexity. This aligns with earlier theory proposing that writing ability is not monolithic but genre-dependent (Hyland, 2007). In addition, pedagogic context—i.e., ESL or EFL context—required an emphasis on certain measures. Zhang & Kang (2022), for instance, reported that test-focused EFL learners produced formulaic forms more than ESL learners pointed to how pedagogic contexts mediate language acquisition. These findings support a modification toward contextually adaptive assessment frameworks, as rigid application of fixed metrics may obscure nuanced learner differences.

RQ 5: Challenges and limitations in measure utilization

The results showed a pattern of persistent issues with the application of fine-grained linguistic metrics in L2 writing assessments. Among the limitations is the fact that there are bounds to computerized measures' accuracy and interpretative trustworthiness. For instance, Peng et al. (2023) found that while lexical complexity and range were associated with more proficient learners' writing scores, more conventional surface-level measures like sentence length more strongly forecasted human raters' scores. This has implications for computer output and human judgment incongruence with implications for computer-mediated test validity. Similarly, Chau (2023) identified syntactic and lexical pattern differences across L1 groups, undermining the cross-linguistic generalizability of such measures as L2SCA and TAALES, which are English data-heavy. Cross-examining the findings through the lens of the tools utilized, entails that genre and proficiency differences only aggravate the measurement reliability challenge. Alshehri (2024) and Zhang & Liu (2021) depicted that syntactic complexity was a better predictor in argumentative compared to narrative writing, with cohesion and rhetorical purposes having more influential roles in the latter. Access is still an issue; technical competence or English-language interfaces are required by most of the tools, such that they are hindered in adoption in low-resource settings (Hassanzadeh et al., 2021; Monteiro et al., 2020). Contrary to earlier theories advocating holistic scoring (Hyland, 2007), contemporary studies concur with the move towards fine-grained examination while upholding the significance of making use of validated context-sensitive measurements. As much as technology has streamlined measurement accuracy, limits of interpretability,

A Systematic Review of Fine-Grained Measures used in L2 Writing Assessment Research

adaptability limitations, and access problems lean towards these instruments being substitutive for expert human judgment rather than adding to it.

5.1 Implications for research and practice

The findings underscore critical implications for both research and practice in L2 writing assessment. Researchers with a set of fine-grained measures must consider the correspondence between language measures and genre-typical, proficiency-sensitive requirements on L2 writing tests. For example, lexical sophistication will prove more helpful in argumentative writing, whereas more narrative genres are helped by measures of cohesion and discourse organization (Alshehri, 2024; Zhang & Liu, 2021). Accordingly, researchers are recommended to become attuned to tasks when choosing analytic measures to steer clear of blanket application. Additionally, the degree of variation in computational instruments employed—ranging from TAALES to Coh-Metrix—facilitates strong demands for tool standardization and cross-validation to ensure consistency and replicability in research (Chau, 2023; Peng et al., 2023). Standardization would also enable comparative research across linguistic and instructional environments. The absence of cross-linguistic calibration in such measures as L2SCA or Coh-Metrix restricts their use to multilingual samples (Khushik & Huhta, 2020). The future research direction should therefore be the creation of validated, multilingual corpora and flexible measures. Moreover, putting together the measures—i.e., syntactic, lexical, and cohesion measures—into an aggregated evaluation framework would represent a broader perspective of writing proficiency (Zhang & Lu, 2022; Heng et al., 2023). Such a multi-perspective method, based on machine analysis as well as human intuition, can lead to more educationally sound and justifiable evaluative practice.

5.2 Limitations of the review

Several limitations delimit the generalizability and quality of the present review of fine-grained measures in L2 writing testing. Among the central issues is publication bias, in that the review included mainly peer-reviewed journal articles and dissertations, which can overrepresent research studies with important or positive results but underreport null or inconclusive results (Chau, 2023; Peng et al., 2023). This bias may misrepresent the true effectiveness and prevalence of linguistic interventions. Furthermore, the stated exclusion of non-English research restricts the linguistic and cultural range of research literature. As Khushik and Huhta (2020) and Chau (2023) noted, syntactic and lexical characteristics tend to differ significantly across L1 groups, and so findings derived from English-language articles only may not be cross-linguistically valid. In addition, inconsistency in reporting methodological information across the studies impeded comparison and synthesis. Calculation software used or indices employed in some of the studies were not reported (Hassanzadeh et al., 2021; Alshehri, 2024), while in others, rating scales or validation processes were not explained. Heterogeneity bars transparency and reproducibility, which are also necessary to determine the validity of results. Overall, these are flaws that call for tighter rates of inclusiveness, more emphasis on methodology, and efforts at minimizing publication bias in systematic reviews and meta-analyses of literature in the future.

6. CONCLUSION

This systematic review pooled data from 52 empirical studies of fine-grained L2 writing measures of assessment and reported prime trends, methodological trends, and underrepresentation trends. The results discovered that lexical and syntactic measures—specifically lexical sophistication, MTLT, and T-unit-based syntactic complexity—were the most typical measures with replicable correlations with writing quality in academic and argumentative genres (Peng et al., 2023; Alshehri, 2024). Such features have been operationally applied by using tools like Coh-Metrix and L2SCA, though issues of their cross-context validity are still prevalent (Chau, 2023; Zhang & Lu, 2022). By contrast, cohesion features and more sophisticated discourse features such as argument structure and organization by genre are not leveraged to their full potential even though they can be applied to narrative and reflective text (Zhang & Liu, 2021; Thi et al., 2023). Because single indices often miss genre and proficiency differences, multiple measures are required to accurately reflect the complexity of L2 writing. It is proposed for future work that cross-linguistic validation of computational tools must be carried out, in addition to consideration of underrepresented learner groups like speakers of typologically dissimilar languages and students enrolled in non-English instructional programs (Khushik & Huhta, 2020). It would enhance generalizability in addition to offering insight into the construction of fair and contextualized assessment frameworks. These are some of the aspects that are required for augmenting theoretical understanding and pedagogic approaches to L2 writing assessment.

Data Availability: The data supporting the conclusions of this systematic review are included within the manuscript and its supplementary material. Further data are available from the author upon reasonable request.

Funding: This research is not funded.

Conflicts of Interest: The author declares no conflict of interest.

REFERENCES

- 1) Alaei, F., Kaivanpanah, S., & Mohammadi, M. (2024). Syntactic complexity and lexical diversity in L1/L2 writing of EFL learners. *Journal of English Language Teaching and Learning*, 16(33), 97-115. <https://doi.org/10.22034/elt.2024.59438.2590>
- 2) Alkhalaf, S. A. A. (2022). Expressive writing in a Saudi university English foreign language (EFL) classroom: Evaluating gains in syntactic complexity. *F1000Research*, 11, 723. <https://doi.org/10.12688/f1000research.121577.1>
- 3) Alsahlanee, A., & Jaganathan, P. (2022). Lexical complexity in the writings of Iraqi, English L2, and English L1 writers. *Asian Journal of University Education*, 18(4), 1105-1118. <http://dx.doi.org/10.24191/ajue.v18i4.20044>
- 4) Alshehri, A. (2022). Writing proficiency of Saudi EFL learners: Examining the impact of lexical diversity. *Journal of Language and Linguistic Studies*, 18(Special Issue 1), 519-529.
- 5) Alshehri, N. (2024). The role of task type in enhancing writing fluency: A study of Saudi female university students. *International Journal of Linguistics, Literature and Translation*, 7(10), 1-7. <https://doi.org/10.32996/ijlilt.2024.7.10.1>
- 6) Aminovna, B. D. (2022). Importance of coherence and cohesion in writing. *Eurasian Research Bulletin*, 4, 83-89. <https://geniusjournals.org/index.php/erb/article/view/431>
- 7) Ansarin, A. A., Karafkan, M. A., & Hadidi, Y. (2021). The effects of task type on Iranian EFL learners' use of lexical diversity and sophistication. *Applied Research on English Language*, 10(4), 39-70. <https://doi.org/10.22108/are.2021.126660.1673>
- 8) Ayadi, M. (2023). Lexical richness and syntactic complexity as predictors of academic writing performance. *Journal of English Studies in Arabia Felix*, 2(1), 23-33. <https://doi.org/10.56540/jesaf.v2i1.43>
- 9) Bae, J., & Min, S. (2020). Genre-based analysis of syntactic complexity in L2 college students' writing: pedagogic scope and directions. *Journal of Asia TEFL*, 17(3), 937. <http://dx.doi.org/10.18823/asiatefl.2020.17.3.12.937>
- 10) Bi, P., & Jiang, J. (2020). Syntactic complexity in assessing young adolescent EFL learners' writings: Syntactic elaboration and diversity. *System*, 91, 102248. <https://doi.org/10.1016/j.system.2020.102248>
- 11) Bulté, B., & Housen, A. (2014). Conceptualizing and measuring short-term L2 grammatical development. *Journal of Second Language Writing*, 26, 42-65. <https://doi.org/10.1016/j.jslw.2014.09.005>
- 12) Chau, T. C. (2023). *The relationships between L1, writing quality, and complexity, accuracy, and fluency in L2 writing* [Doctoral dissertation, University of South Florida]. <https://digitalcommons.usf.edu/etd/10106>
- 13) Cheung, Y.L., Chu, Y., & Jang, H. (2021). Impact of a socio-cognitive approach to teaching English language writing on primary school students' compositions. *Iranian Journal of Language Teaching Research*, 9(1), (Jan., 2021) 1-22. <https://doi.org/10.30466/ijltr.2021.120973>
- 14) Chomsky, N. (1995). *The minimalist program*. MIT Press.
- 15) Chon, Y. V., & Shin, D. (2020). Direct writing, translated writing, and machine-translated writing: A text level analysis with Coh-metrix. *English Teaching*, 75(1), 25-48. <https://doi.org/10.15858/engtea.75.1.202003.25>
- 16) Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The development and use of cohesive devices in L2 writing and their relation to judgments of essay quality. *Journal of Second Language Writing*, 32, 1-16. <https://doi.org/10.1016/j.jslw.2016.01.003>
- 17) Esfandiari, R., & Ahmadi, M. (2024). Syntactic complexity in novice and expert L2 academic writing: The role of syntactic elaboration and diversity. *Journal of Language Horizons*, 8(1), 115-138. <https://doi.org/10.22051/lghor.2023.43081.1780>
- 18) Golparvar, S. E., & Rashidi, F. (2021). The effect of task complexity on integrated writing performance: The case of multiple-text source-based writing. *System*, 99, 102524. <https://doi.org/10.1016/j.system.2021.102524>
- 19) Graesser, A. C., McNamara, D. S., McCarthy, P. M., & Cai, Z. (2014). *Coh-Metrix: An automated tool for theoretical and applied natural language processing*. Routledge.
- 20) Granados, A., López-Jiménez, M. D., & Lorenzo, F. (2022). A longitudinal study of L2 historical writing. *Ibérica*, 43, 129-154. <https://doi.org/10.17398/2340-2784.43.129>
- 21) Hassanzadeh, M., Saffari, E., & Rezaei, S. (2021). The impact of computer-aided concept mapping on EFL learners' lexical diversity: A process writing experiment. *ReCALL*, 33(3), 214-228. <https://doi.org/10.1017/S095834402100001X>
- 22) He, Z. (2020). Cohesion in academic writing: a comparison of essays in English written by L1 and L2 university students. *Theory & Practice in Language Studies*, 10(7), 761-770. <http://dx.doi.org/10.17507/tpls.1007.06>
- 23) Heng, R., Pu, L., & Liu, X. (2023). The effects of genre on the lexical richness of argumentative and expository writing by Chinese EFL learners. *Frontiers in Psychology*, 13, 1082228. <http://dx.doi.org/10.17507/tpls.1007.06>
- 24) Hyland, K. (2007). Genre pedagogy: Language, literacy and L2 writing instruction. *Journal of second language writing*, 16(3), 148-164. <https://doi.org/10.1016/j.jslw.2007.07.005>
- 25) Istiharoh, A. F., & Pranowo, D. D. (2023). The development of syntactic complexity in university students' essays. *Englisia: Journal of Language, Education, and Humanities*, 11(1), 1-18. <https://doi.org/10.22373/ej.v11i1.19129>

A Systematic Review of Fine-Grained Measures used in L2 Writing Assessment Research

- 26) Karadeniz, A. (2017). Cohesion and coherence in written texts of students of faculty of education. *Journal of Education and Training Studies*, 5(2), 93-99. <http://dx.doi.org/10.11114/jets.v5i2.1998>
- 27) Kessler, M., Ma, W., & Solheim, I. (2022). The effects of topic familiarity on text quality, complexity, accuracy, and fluency: A conceptual replication. *TESOL Quarterly*, 56(4), 1163-1190. <https://doi.org/10.1002/tesq.3096>
- 28) Khushik, G. A., & Huhta, A. (2020). Investigating syntactic complexity in EFL learners' writing across common European framework of reference levels A1, A2, and B1. *Applied Linguistics*, 41(4), 506-532. <https://doi.org/10.1093/applin/amy064>
- 29) Kim, M. (2021). Exploring longitudinal changes in lexical and syntactic features in beginning level EFL learner writing. *System*, 103, 102680. <https://doi.org/10.1016/j.system.2021.102680>
- 30) Kim, M., Kim, Y., & Kang, S. (2024). The dynamics of changes in linguistic complexity and writing scores in timed argumentative writing among beginning-level EFL learners. *International Review of Applied Linguistics in Language Teaching*, 63(4):2891-2925. <https://doi.org/10.1515/iral-2023-0287>
- 31) Krashen, S. D. (1985). *The input hypothesis: Issues and implications*. Longman.
- 32) Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, 49(4), 757-786. <https://doi.org/10.1002/tesq.194>
- 33) Kyle, K., & Crossley, S. A. (2018). Assessing syntactic sophistication in L2 writing: A usage based approach. *Language Testing*, 35(4), 515-535. <https://doi.org/10.1177/0265532217712554>
- 34) Kyle, K., Crossley, S., & Verspoor, M. (2021). Measuring longitudinal writing development using indices of syntactic complexity and sophistication. *Studies in Second Language Acquisition*, 43(4), 781-812. <https://doi.org/10.1017/S0272263120000546>
- 35) Lahuerta Martínez, A.C. (2024). The dynamics of syntactic and lexical complexity measures in academic writing. *Ibérica*, 251-274. <https://doi.org/10.17398/2340-2784.47.251>
- 36) Lahuerta, A. C. (2023). L2 Writing: A comparison of upper intermediate and advanced EFL learners. *European Journal of Language and Literature Studies*, 9(2), 64-71. <https://doi.org/10.2478/ejls-2023-0016>
- 37) Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. MIT Press.
- 38) Li, Y., Nikitina, L., & Riget, P. N. (2022). Development of syntactic complexity in Chinese university students' L2 argumentative writing. *Journal of English for Academic Purposes*, 56, 101099. <https://doi.org/10.1016/j.jeap.2022.101099>
- 39) Lipková M. (2024). Lexical density in academic writing: Lexical features and learner corpora analysis in L2 tertiary students' essays and didactic implications. *The Asian Conference on Education 2023: Official Conference Proceedings* (pp. 1175-1188) <https://doi.org/10.22492/issn.2186-5892.2024.98>
- 40) Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15(4), 474-496. <https://doi.org/10.1075/ijcl.15.4.02lu>
- 41) Malvern, D., Richards, B., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development: Quantification and assessment*. Palgrave Macmillan.
- 42) McCarthy, P. M., & Jarvis, S. (2010). MTL-D, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42(2), 381-392. <https://doi.org/10.3758/BRM.42.2.381>
- 43) McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic features of writing quality. *Written communication*, 27(1), 57-86. <https://doi.org/10.1177/0741088309351547>
- 44) McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press.
- 45) Monteiro, K. R., Crossley, S. A., & Kyle, K. (2020). In search of new benchmarks: Using L2 lexical frequency and contextual diversity indices to assess second language writing. *Applied Linguistics*, 41(2), 280-300. <https://doi.org/10.1093/applin/amy056>
- 46) Nation, I. S. P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge University Press. <https://doi.org/10.1017/CBO9781139858656>
- 47) Nguyen, B. T. T., Newton, J., & Tran, Q. P. N. (2022). The effect of topic on EFL writing by Vietnamese tertiary students: Insights from combining a lexical richness analysis with student self-reports. *Language Related Research*, 13(5), 511-540. <http://dx.doi.org/10.52547/LRR.13.5.19>
- 48) Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics*, 24(4), 492-518. <https://doi.org/10.1093/applin/24.4.492>
- 49) Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., . . . Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, Article n71. <https://doi.org/10.1136/bmj.n71>

A Systematic Review of Fine-Grained Measures used in L2 Writing Assessment Research

- 50) Peng, Y., Sun, J., Quan, J., Wang, Y., Lv, C., & Zhang, H. (2023). Predicting Chinese EFL learners' human-rated writing quality in argumentative writing through multidimensional computational indices of lexical complexity. *Assessing Writing*, 56, 100722. <https://doi.org/10.1016/j.asw.2023.100722>
- 51) Peng, Y., Zheng, Y., Sun, J., Jiang, Y., Lin, J., & Zhang, H. (2024). Modeling relationships among large-grained, fine-grained absolute syntactic complexity and assessed L2 writing quality: An SEM approach. *Assessing Writing*, 61, 100875. <https://doi.org/10.1016/j.asw.2024.100875>
- 52) Pienemann, M. (2005). *Cross-linguistic aspects of processability theory*. John Benjamins.
- 53) Qian, L., Cheng, Y., & Zhao, Y. (2021). Use of linguistic complexity in writing among Chinese EFL learners in high-stakes tests: Insights from a corpus of TOEFL iBT. *Frontiers in Psychology*, 12, 765983. <https://doi.org/10.3389/fpsyg.2021.765983>
- 54) Qin, W., & Uccelli, P. (2020). Beyond linguistic complexity: Assessing register flexibility in EFL writing across contexts. *Assessing Writing*, 45, 100465. <https://doi.org/10.1016/j.asw.2020.100465>
- 55) Rivard, L. P., & Gueye, N. R. (2023). Syntactic complexity and connector use in the summary writing of L1 and L2 Canadian students. *Journal of French Language Studies*, 33(2), 197-226. <https://doi.org/10.1017/S0959269522000291>
- 56) Seo, J. Y. (2024). Exploring the educational potential of ChatGPT: AI-assisted narrative writing for EFL college students. *Language Teaching Research Quarterly*, 43, 1-21. <https://doi.org/10.32038/ltrq.2024.43.01>
- 57) Susoy, Z. & Köse, D. G. (2024). An evaluation of syntactic complexity, lexical diversity and text length as predictors of foreign language writing scores. *Erzincan University Journal of Education Faculty*, 26(4), 542-553. <https://doi.org/10.17556/erziefd.1483282>
- 58) Sweller, J. (2011). *Cognitive load theory*. *Psychology of Learning and Motivation*, 55, 37-76. <https://doi.org/10.1016/B978-0-12-387691-1.00002-8>
- 59) Tabari, M. A., & Wind, A. (2023). Dynamic development of cohesive devices in English as a second language writing. *International Review of Applied Linguistics in Language Teaching*, 63(1), 433-469. <https://doi.org/10.1515/iral-2022-0205>
- 60) Tabari, M. A., Lu, X., & Wang, Y. (2023). The effects of task complexity on lexical complexity in L2 writing: An exploratory study. *System*, 114, 103021. <https://doi.org/10.1016/j.system.2023.103021>
- 61) Thi, N. K., Vo, D. V., & Nikolov, M. (2023). Investigating syntactic complexity and language related error patterns in EFL students' writing: corpus-based and epistemic network analyses. *Language Learning in Higher Education*, 13(1), 127-151. <https://doi.org/10.1515/cercles-2023-2009>
- 62) Tian, Y., Kim, M., Crossley, S., & Wan, Q. (2021). Cohesive devices as an indicator of L2 students' writing fluency. *Reading and Writing*, 1-23. <https://doi.org/10.1007/s11145-021-10229-3>
- 63) Ure, J. (1971). Lexical density and register differentiation. In G. E. Perren & J. L. M. Trim (Eds.), *Applications of linguistics* (pp. 443-452). Cambridge University Press.
- 64) Wang, Y. (2023). The interplay of linguistic complexity measures and their relation to the quality of L2 argumentative writing. *Proceedings of the 4th International Conference on Educational Innovation and Philosophical Inquiries*. <https://doi.org/10.54254/2753-7064/11/20231365>
- 65) Weigle, S. C. (2002). *Assessing writing*. Cambridge University Press.
- 66) Wind, A. M. (2022). The development of syntactic complexity and fluency in an advanced writing course. In A. Wind & J. Csizér (Eds.), *DEAL 2022: Challenges and Opportunities in Contemporary English Applied Linguistics* (pp. 97-118). Eötvös University Press. <https://doi.org/10.21862/ELTE.DEAL.2022.4.97>
- 67) Wondemagegne, E., Simegn, B., & Adugna, E. T. (2024). Exploring syntactic complexity and its relationship with writing quality in EFL argumentative essays. *Journal of Language Research*, 1 (2), 13-35. <https://doi.org/10.14302/issn.2998-4122.jlr-24-4985>
- 68) Xue, Y. (2024). Towards automated writing evaluation: A comprehensive review with bibliometric, scientometric, and meta-analytic approaches. *Education and Information Technologies*, 29(15), 19553-19594. <http://dx.doi.org/10.1007/s10639-024-12596-0>
- 69) Yang, Y., Yap, N. T., & Ali, A. M. (2023). Predicting EFL expository writing quality with measures of lexical richness. *Assessing Writing*, 57, 100762. <https://doi.org/10.1016/j.asw.2023.100762>
- 70) Zenouzagh, Z. M. (2020). Syntactic complexity in individual, collaborative and E collaborative EFL writing: mediating role of writing modality, L1 and sustained development in focus. *Educational Technology Research and Development*, 68(6), 2939-2970. <https://doi.org/10.1007/s11423-020-09818-w>
- 71) Zhang, C., & Kang, S. (2022). A comparative study on lexical and syntactic features of ESL versus EFL learners' writing. *Frontiers in Psychology*, 13, 1002090. <https://doi.org/10.3389/fpsyg.2022.1002090>
- 72) Zhang, X., & Lu, X. (2022). Revisiting the predictive power of traditional vs. fine-grained

A Systematic Review of Fine-Grained Measures used in L2 Writing Assessment Research

- syntactic complexity indices for L2 writing quality: The case of two genres. *Assessing Writing*, 51, 100597. <https://doi.org/10.1016/j.asw.2021.100597>
- 73) Zhang, C. (2022). Multidimensional interaction of L2 writing fluency and cohesion: A complexity theory perspective. *Southern African Linguistics and Applied Language Studies*, 40(4), 397-412. <https://doi.org/10.2989/16073614.2022.2075773>
- 74) Zhang, L. & Liu, H. (2021). Genre effect on L2 syntactic complexity and holistic rating for writing quality of intermediate EFL learners. *Chinese Journal of Applied Linguistics*, 44(4), 451-469. <https://doi.org/10.1515/CJAL-2021-0029>
- 75) Zhang, X., & Li, W. (2024). Remodeling effects of linguistic features on L2 writing quality of two genres. *Reading and Writing*, 37(5), 1209-1234. <https://doi.org/10.1007/s11145-023-10438-y>
- 76) Zhang, J. & Zhang, L. (2024). Examining L2 students' development of global cohesion and its relationship with their argumentative essay quality. *International Review of Applied Linguistics in Language Teaching*. <https://doi.org/10.1515/iral-2023-0069>
- 77) Zindela, N. (2023). Comparing measures of syntactic and lexical complexity in artificial intelligence and L2 human-generated argumentative essays. *International Journal of Education and Development Using Information and Communication Technology*, 19(3), 50-68. <https://files.eric.ed.gov/fulltext/EJ1413411.pdf>