

Original Article



Conceptualization of Complexity, Accuracy, and Fluency in Assessing Medical Writings: Comparing language and Content Specialist

Sayyed Mohammad Alavi^{1*}, Neda Rahmatvand²¹Faculty of Foreign Language and Literatures, University of Tehran, Tehran, Iran²Faculty of Foreign Language and Literatures, Alborz Campus, University of Tehran, Tehran, Iran

Article info

Article History:

Received: October 4, 2025

Revised: December 1, 2025

Accepted: December 1, 2025

Published: August 2, 2026

Keywords:

Complexity, Accuracy, Fluency,
CAF framework, Medical writing

Abstract

Introduction: This study investigates how language specialists and content experts conceptualize and prioritize the dimensions of Complexity, Accuracy, and Fluency (CAF) in the assessment of medical writing.**Methods:** Six hypertension-related medical texts were independently evaluated by medical specialists and language specialists using a standardized CAF rubric.**Results:** The quantitative analysis showed near-perfect inter-rater agreement for Accuracy, substantial agreement for Complexity, and moderate agreement for Fluency. A qualitative thematic analysis of raters' justifications and panel discussions revealed a shared emphasis on the critical importance of accuracy, nuanced differences in the interpretation of complexity, and divergent views on fluency shaped by stylistic and disciplinary preferences. The findings highlight the central tension between maintaining clinical precision and ensuring language accessibility, underscoring the importance of interdisciplinary evaluation for comprehensive assessment of medical texts.**Conclusion:** This integrated approach offers a replicable model for evaluating medical texts that balances clinical accuracy with linguistic clarity, ultimately supporting clear and effective scientific communication. This study contributes valuable insights to language assessment and medical writing pedagogy, supporting tailored interventions to improve the linguistic quality and practical effectiveness of scientific texts authored by non-native English speakers.

Introduction

The Complexity, Accuracy, and Fluency (CAF) framework, introduced by Skehan in 1989 as a three-dimensional model of second language proficiency, has been examined from various theoretical perspectives across diverse languages and proficiency levels.¹⁻⁴ In many of these studies, CAF has been studied as a dependent variable for various tasks and genres on oral and written performance (for example,^{5,6}). Previous research has explored the developmental trajectories of CAF during the language acquisition process. A central focus of such studies is examining the ways in which CAF evolve over time.^{7,8} An important concern of this type of research is how CAF develops over time. So far, most studies on how CAF develops in learning a second language have mainly tested specific ideas. Building on the historical evolution of the CAF framework, we specifically examine its application within the specialized context of medical writing. This focus underscores the framework's

adaptability and relevance across different fields, setting the stage for the present study.

To complement the findings of previous research, the goal of the present study is to explore how language specialists and medical experts perceive and apply CAF in evaluating medical texts. By clarifying how these two groups conceptualize CAF, the research reveals how assessor beliefs and backgrounds shape and operationalize CAF dimensions, sources of scoring inconsistencies between linguistic and domain priorities, and highlights the implications identified of these assessment practices. The findings will offer evidence-based recommendations for training assessors, delivering feedback informed by CAF, and developing university editorial programs. This will support fairer, more diagnostic, and discipline-sensitive assessments in EFL medical education. Ultimately, this study holds practical value in enhancing both the linguistic quality and disciplinary accuracy of medical manuscripts written by L2 authors. This will guide

*Corresponding author: Sayyed Mohammad Alavi, Email: smalavi@ut.ac.ir

targeted professional development, editorial assistance, and assessment strategies to improve manuscript quality and promote more effective dissemination of medical research.

Building on the discussion of CAF's importance in medical writing, we delve into how teachers' perceptions of CAF may vary based on syntactic features, subject preferences, and pedagogical contexts, this study compares how teachers from two distinct academic backgrounds—language specialists and medical specialists—perceive CAF. To understand how different academic settings influence the perception of CAF, we conducted the study at two institutions: Tabriz and Tehran University of Medical Sciences. This approach allows us to compare and contrast the impact of varying academic environments on CAF assessment. Two groups of teachers, one comprising language specialists and the other medical specialists, were asked to evaluate samples of argumentative texts written by medical university students and provide feedback. In a subsequent panel discussion, the teachers explained the reasoning behind their scores and elaborated on the feedback they offered to the students.

In recent years, AI-driven medical text evaluation has become an emerging focus in clinical informatics and medical education research, reflecting a growing need for reliable frameworks to assess the quality and coherence of both human- and AI-generated medical discourse. Studies such as Croxford, et al.,⁹ have introduced LLM-as-a-Judge systems that use large language models to evaluate medical summaries, demonstrating promising consistency with human judgments but raising ongoing concerns about methodological transparency and reproducibility. Similarly, systematic reviews of AI-based clinical documentation and speech recognition tools¹⁰ have revealed significant variation in evaluation criteria and limited reporting of data validity. Broader analyses of AI evaluation frameworks in medicine¹¹ highlight a lack of standardized methodologies and the need for clearer operationalization of assessment rubrics. Parallel research in AI applications for medical education¹² and the evaluation of generative AI tools in interpreting medical texts¹³ further emphasize the importance of well-defined, transparent evaluation procedures. Despite these advancements, current approaches remain fragmented, underscoring the necessity for an integrated, theoretically grounded framework that explicitly defines assessment constructs and analytical procedures. The present study addresses this gap by employing a CAF-based evaluation model to provide a transparent, replicable method for analyzing medical text performance.

The rationale for this study is to help bridge the gap between research and classroom practice, grounded in the belief that teachers and researchers can mutually benefit from one another. On one hand, theoretical frameworks and research findings offer teachers deeper insights into second language (L2) learning and their specific teaching

contexts, encouraging them to consider alternative pedagogical strategies. On the other hand, teachers' reflections on the successes and challenges they encounter in the classroom provide researchers with valuable, real-world perspectives on language learning, often referred to as learning in authentic contexts. These insights can inspire new research questions and directions.¹⁴

The following sections present the study's background through a review of CAF in academic writing assessment. Next, the paper discusses how writing is measured using the CAF framework and examines assessor perceptions of CAF in written performance. The methodology section then outlines the study's design. This is followed by a comparison of the two groups of teachers regarding their rating motivations and feedback practices. In the final section, the findings are summarized and interpreted in light of SLA theory and CAF research, followed by a discussion of their pedagogical implications for classroom practice and teacher training.

CAF in Academic Writing Assessment

Academic writing is a complex process that involves multiple cognitive and linguistic dimensions, including mastery of academic vocabulary, grammatical accuracy, and syntactic sophistication—all of which contribute to the production of coherent and precise texts.¹⁵ It also requires critical thinking and the strategic integration of multiple sources, enabling writers to evaluate evidence and incorporate it effectively to substantiate their arguments. Scientific writing, in particular, serves as a key means of academic communication, allowing researchers to share their findings and contribute to the expansion of global knowledge. Academic writing thus represents a structured, evidence-based form of discourse that enables scholars to disseminate research findings within established disciplinary conventions. It is characterized by clarity, conciseness, and coherence, ensuring that arguments are logically developed and supported by credible sources.¹⁶

Effective communication is especially vital in medical research, where it serves as a primary medium for disseminating knowledge, facilitating cross-disciplinary collaboration, and promoting innovation in healthcare. In this rapidly evolving field, researchers must convey complex medical concepts and research outcomes with precision and clarity.¹⁷ Medical students, in particular, face distinctive challenges in academic writing, as they are required to master specialized terminology, employ language with accuracy, and adhere to the formal conventions of scientific discourse.¹⁸

Evaluating students' writing and providing feedback are fundamental practices in English language writing instruction. Feedback from teachers significantly influences students' development as English writers. As Casanave¹⁹ observes, "*Many of the decisions that both L1 and L2 writing teachers make in their classes revolve*

around assessment of students' writing... (and) because a culture of assessment is built into the schooling enterprise, teachers rarely ask whether they need to assess their students" (p. 113). Despite this, it is essential for teachers to reflect on how and why they provide feedback, as such reflection forms an important part of professional growth and enhances their effectiveness as educators. Rubrics are particularly useful tools in language classrooms, as they allow teachers to evaluate multiple components of written assignments, including rhetorical structure, grammatical accuracy, and topic development. They also encourage teachers and evaluators to focus on more than just sentence-level structures, enabling a more holistic assessment of student writing.

The Complexity, Accuracy, and Fluency (CAF) framework originated in the 1970s and 1980s, during efforts to establish quantitative and objective measures for assessing second language (L2) proficiency and development. A pioneering study by Larsen-Freeman²⁰ emphasized the need for valid and reliable indicators of L2 proficiency, marking the beginning of CAF as a comprehensive framework for evaluating linguistic performance. The framework captures three key dimensions of second language writing proficiency. Complexity, often regarded as the "*most complex, ambiguous, and least understood dimension of the CAF triad*"²¹, refers to the degree of elaboration and structural sophistication in language use, reflecting a learner's ability to employ diverse vocabulary and advanced syntactic constructions. Complexity encompasses several subdimensions of language use, including lexical, interactional, propositional, and grammatical complexity.²²⁻²⁴

Accuracy is defined as the extent to which language production conforms to the norms of the target language, focusing on the correct and error-free use of grammar and vocabulary. This dimension is particularly critical in formal and academic contexts, where linguistic precision is essential.^{25,26}

Fluency, the third dimension, is described as an "*indicator of comfort in producing language or the ease of language retrieval*".²⁷ It reflects the smoothness and efficiency of language production, demonstrating a speaker's ability to communicate with minimal cognitive strain or linguistic disruption.³

Together, these three dimensions provide a comprehensive framework for assessing L2 proficiency, allowing educators and researchers to identify learners' strengths and weaknesses across multiple aspects of linguistic competence and to design targeted instructional interventions. Larsen-Freeman²⁸ and Skehan²⁹ emphasize that effective second-language writing proficiency requires the simultaneous development of all three dimensions rather than isolated improvement in any single one.

Academic writing represents a unique linguistic style²² that poses considerable challenges for L2 learners. While

it is recognized as a highly complex and sophisticated form of language, Biber and Gray²² demonstrated that its complexity does not arise from longer sentences or greater clausal subordination typical of speech. Instead, academic writing is more "*compressed*", relying heavily on phrasal, non-clausal modifiers embedded within noun phrases (p. 7). Thus, the intricate nature of academic writing—contrasting with spoken discourse—lies primarily in its preference for embedded phrases rather than extensive use of subordinate clauses. Therefore, they hypothesized that L2 learners develop their use of embedded phrases in academic writing through progressive stages, with each stage showing greater syntactic elaboration than the last. Parkinson and Musgrave³⁰ found a positive correlation between learners' proficiency levels and the frequency of noun modifiers in their academic texts. Similarly, Crossley and McNamara³¹ documented an increase in both noun usage and phrasal complexity in the writing of L2 students over the course of a semester.

Beyond syntactic characteristics, academic writing is also distinguished by its vocabulary profile. The use of academic vocabulary is closely tied to a writer's overall competence, functioning not only as lexical choice but also as a tool for constructing arguments, signaling logical relationships, and expressing degrees of certainty or knowledge. Coxhead³² conducted an in-depth analysis to identify words that frequently appear in academic texts but are not included in West's General Service List. This research led to the development of the Academic Word List (AWL), which contains 570 word families and accounts for about 10% of the vocabulary used in academic writing—a much higher proportion than the 1.4% found in fiction texts. Together, the AWL and the General Service List cover approximately 86.1% of the words in academic texts, and Douglas³³ later estimated that this coverage could reach up to 94% in essays written by novice native English-speaking university students. Kong³⁴ further noted that the ability to use academic vocabulary effectively reflects a writer's capacity to organize ideas, think critically, and produce clear, well-structured academic texts.

Taken together, these linguistic and lexical features underscore the inherent complexity of academic writing and its close alignment with the CAF framework, in which syntactic elaboration, lexical sophistication, and accuracy interact to shape the overall proficiency of L2 writers.

Measuring CAF

Complexity in second language (L2) writing is widely regarded as a critical indicator of linguistic proficiency, reflecting the degree of linguistic sophistication and elaboration present in a learner's language production. This construct is typically operationalized at both the syntactic and lexical levels. Syntactic complexity pertains to the structural intricacy of sentences and is frequently measured using indices such as mean length of terminable

units or clauses, subordination ratio—defined as the frequency of dependent clauses relative to total clauses—complex nominals per clause, and coordinate phrases per clause. Collectively, these measures capture the extent of syntactic elaboration by assessing dependent clause embedding, noun phrase complexity, and coordination strategies. Such indices serve as valuable proxies for tracking developmental progress in acquiring more advanced syntactic structures. Conversely, lexical complexity refers to the richness and diversity of a learner's vocabulary, commonly evaluated through metrics including type-token ratio, lexical diversity, and the frequency of low-frequency lexical items. These indicators collectively denote an expanded and sophisticated lexical repertoire, which is essential for effective academic writing and communicative competence.³⁵

Khushik and Huhta³⁶ studied syntactic complexity in EFL learners' writing at CEFR A1, A2, and B1 levels. They analyzed 28 complexity measures in essays by Pakistani and Finnish teenagers using computational tools. Most measures, especially length of production units, subordination, and phrasal density, effectively distinguished proficiency levels. Differences between the two L1 groups appeared mainly in length and phrasal complexity. The study supports syntactic complexity as a key indicator of CEFR-based writing development.

The methodologies employed to assess syntactic complexity in second language writing have been extensively studied, with several researchers contributing to a complex understanding of this construct. Lu's³⁷ classification of syntactic complexity measures provides a comprehensive overview of the commonly used indices that capture the multidimensional nature of syntactic complexity. The first category encompasses measures based on the length of linguistic production, including mean length of clause, mean length of sentence, and mean length of terminable unit. The second category focuses on sentence complexity, represented by the measure of clauses per sentence. The third category consists of four measures that assess subordination within the text, evaluating the extent of dependent clause usage. The fourth category includes measures that quantify the degree of coordination present in the text. Together, these categories offer a systematic framework for analyzing different dimensions of syntactic complexity.

Lexical complexity in second language writing is defined by three fundamental dimensions: lexical density, lexical diversity, and lexical sophistication. Lexical density quantifies the proportion of content words within a text, whereas lexical diversity reflects the range of vocabulary employed. Lexical sophistication pertains to the utilization of less frequent or more advanced lexical items, often assessed through reference corpus frequencies. Collectively, these dimensions form a robust framework for the systematic evaluation of lexical richness and complexity in second language written production.³⁸

Accuracy in language assessment, like syntactic complexity, is typically measured through norm-based evaluations of grammar and certain lexical errors, excluding semantic or pragmatic aspects.² It involves both global and specific aspects of grammatical and lexical correctness, which are essential for maintaining clarity and credibility in written communication. The methodologies employed to assess accuracy in second language writing typically involve a combination of quantitative and qualitative approaches. Quantitative measures often include calculating the proportion of error-free units, such as error-free clauses or T-units, relative to the total number of units produced.³⁹ Pallotti²⁴ defines accuracy as the degree of adherence to norms, emphasizing its link to target-like language use. Unlike complexity, accuracy reflects language control rather than flexibility. In writing, accuracy is assessed by counting errors, calculating error-free phrase or clause ratios, holistic or analytic ratings, and weighted error ratios.²⁷ Research consistently shows a positive correlation between grammatical accuracy and language proficiency.⁴⁰ In integrated writing assessments, Cumming et al.⁴¹ and Gebril and Plakans⁴² used holistic scales to compare grammatical accuracy across score levels. They found that lower scoring writings were less accurate, while higher scores correlated with greater accuracy. However, Gebril and Plakans⁴² also noted that the accuracy gap was mostly significant at the lowest levels, with little difference between mid and high-level writings, suggesting accuracy is more distinct among lower proficiency writers. More sophisticated approaches, such as weighted clause ratios, integrate both error frequency and communicative severity to offer a nuanced evaluation of accuracy.⁴³

Fluency is recognized as a critical aspect of writing performance, particularly in second language acquisition and assessment. It refers to the ease and efficiency with which a writer produces text, reflecting both cognitive control and language proficiency. Fluent writing refers to a state of effortless writing characterized by "short pausing times, few revisions and a high production rate".⁴⁴ Writing fluency is shaped by a writer's pausing, revision behaviors, and the rate of text production. Frequent or long pauses, extensive revisions, and slow writing speed all negatively impact fluency. Key features of writing fluency include pausing, revision, and production rate.⁴⁵ It denotes the automaticity and continual flow of language generation, reflecting the writer's capacity to produce language without significant cognitive disruption or interruption.⁴⁶ The measurement of writing fluency typically includes quantitative metrics such as the number of words or characters produced within a given time, the speed of production, and pause or revision frequency during the writing process. These product-based measures are often complemented by process-based measures, which account for the dynamics of writing, including revisions and deletions, thus offering a more precise reflection

of the writer's fluency.⁴⁵ Additionally, automated essay scoring systems assess fluency through features such as vocabulary diversity, sentence structure complexity, and sentence perplexity, leveraging machine learning and natural language processing techniques to evaluate the naturalness and smoothness of writing.⁴⁷ This multifaceted approach situates writing fluency both as a measure of the quantity of written output and its qualitative coherence and linguistic sophistication, both of which are crucial in academic writing.

A comprehensive understanding of how the CAF components manifest in medical manuscripts provides valuable insights for the design and implementation of training programs and editorial interventions aimed at enhancing manuscript quality. The conceptualization of CAF in medical writing closely parallels the challenges and innovations encountered in managing large-scale healthcare data. Accordingly, insights from healthcare data evaluation can inform strategies for assessing medical writing, emphasizing the need for multidimensional assessment tools that balance form, function, and content expertise.⁴⁸

Building on this foundation, the CAF framework serves as an instrumental tool for analyzing and improving writing quality, particularly among researchers who are non-native English speakers. Professional medical writing support often acts as an intermediary, enhancing manuscript quality by strengthening the CAF dimensions to ensure that texts uphold ethical standards and scientific rigor while remaining comprehensible and engaging.⁴⁹ This delicate balance underscores the critical importance of expertise in both subject matter and language use within scientific and medical communication.^{50,51}

Assessor Perceptions of CAF in Writing

Teachers play a central role in both the development and execution of assessment, which is widely recognized as an effective means of enhancing student learning. The effectiveness of assessment largely depends on teachers' perceptions of its value and their ability to implement assessment strategies effectively within classroom contexts.⁵² Teachers' conceptualizations of assessment influence how assessment data are used, guide the design of assessment tools, and shape the extent to which students engage in assessment-related activities. Given the central role of assessment in educational policy, understanding teachers' beliefs about its purposes is essential, as these beliefs directly affect assessment practices.^{53,54} Educators tend to embrace information consistent with their existing beliefs and resist conflicting perspectives, making these conceptions a core component of teacher assessment literacy.⁵⁵

Zheng and Xu⁵⁶ investigated how different teacher assessment approaches affect university students' engagement in writing across various languages. Using a mixed-methods survey, they found that student

engagement has cognitive, behavioral, and affective dimensions influenced by teacher feedback style. Higher proficiency students showed greater cognitive and behavioral engagement, applying more metacognitive strategies and revisions. In contrast, less proficient students had more positive attitudes toward feedback despite lower active engagement. The study highlights the complex relationship between teacher assessment methods and student writing engagement, emphasizing the need for tailored feedback to support diverse learner needs.

Teachers' perceptions of syntactic complexity in L2 writing further illustrate the interplay between theory and practice. Studies on L2-Dutch and L2-Italian teachers show that, despite instructions to focus on syntactic complexity, their evaluations were more strongly guided by grammatical accuracy and clarity. This highlights a gap between theoretical models of CAF and classroom assessment practices, as well as variation influenced by linguistic and contextual factors.⁵⁷ Similarly, research on task types and learner proficiency indicates that fluency and complexity are influenced by writing tasks, whereas accuracy is less affected, underscoring the dynamic interaction of linguistic and cognitive factors in L2 writing.^{4,58,59}

In specialized disciplines such as medicine, the reliability and validity of writing assessments depend on the collaboration between linguistic and content experts. While linguistic assessors focus on grammatical accuracy and lexical appropriateness, content experts emphasize the accuracy, organization, and domain relevance of ideas.⁶⁰ However, inconsistent scoring often arises from these divergent emphases⁶¹, highlighting the need for rater calibration and shared understanding of CAF criteria to ensure fairness and coherence in assessment. The assessment of CAF in medical writing similarly depends on the perceptions and literacy of evaluators, including language specialists and content experts. These assessors often differ in how they prioritize CAF dimensions: language specialists emphasize linguistic correctness and fluency, while content experts focus on factual accuracy and disciplinary relevance.⁶² Effective medical writing assessment therefore requires a systematic, evidence-based approach that ensures clear operationalization of CAF dimensions—complexity, accuracy, and fluency—supported by contextual sensitivity to the demands of medical communication and ongoing assessor training to ensure reliability.⁶³

Professional medical writing expertise also plays a critical role in improving CAF-related outcomes. Systematic reviews indicate that manuscripts developed with professional writing support show higher compliance with reporting standards, clearer English, and stronger ethical and scientific integrity compared to those produced independently.⁶⁴ This highlights the importance of editorial interventions and training in

improving the linguistic and communicative quality of scientific manuscripts.

Despite extensive research on CAF in L2 writing, significant gaps remain in understanding how different assessors conceptualize and apply these dimensions—particularly in specialized genres such as medical writing. Most existing studies have focused on general academic or classroom-based contexts, where assessments are typically conducted by language teachers or researchers with a linguistic background (e.g., ^{3,57}). Consequently, these studies have provided valuable insights into how CAF develops in learner writing but offer limited understanding of how disciplinary or professional expertise shapes evaluators' judgments.

In medical and scientific contexts, writing is evaluated not only for linguistic quality but also for content accuracy, disciplinary conventions, and communicative clarity. This dual focus introduces potential differences in how language specialists and content specialists interpret and prioritize CAF dimensions. While language specialists may focus on syntactic complexity, lexical sophistication, and fluency, content specialists often prioritize factual accuracy, logical organization, and scientific precision.^{60,62} Yet, few empirical studies have directly compared these assessor groups, leaving unclear the extent to which their evaluative judgments converge or diverge.

Furthermore, there is limited understanding of how assessor background, training, and professional orientation influence CAF judgments. Variability in rater interpretations and the absence of standardized calibration across disciplines contribute to subjectivity and inconsistency in writing assessment.^{61,63} This lack of alignment between theoretical models of CAF and practical assessment criteria raises questions about the validity and reliability of evaluations in domain-specific writing, particularly in high-stakes medical contexts.

Another underexplored area concerns the nature and type of feedback provided by assessors from different disciplinary backgrounds. While previous research has shown that feedback plays a central role in shaping students' writing development⁵⁶, there is insufficient evidence regarding how feedback from language versus content specialists differs in focus, tone, and pedagogical intent when applied to CAF-based evaluation.

Taken together, these gaps highlight the need for research that investigates how assessors with different professional orientations conceptualize, operationalize, and apply CAF criteria in medical writing assessment. Addressing this gap will contribute to a more balanced and interdisciplinary understanding of writing assessment, supporting the development of more reliable, context-sensitive, and pedagogically meaningful evaluation frameworks that integrate linguistic proficiency with disciplinary accuracy and communicative effectiveness.

Therefore, there is a pressing need for empirical investigation into how language and content specialists

conceptualize and assess CAF in medical writing. Understanding their differing evaluative priorities and feedback strategies will support the design of balanced, multidimensional assessment frameworks that integrate linguistic proficiency with scientific accuracy and communicative effectiveness.

RQ1: To what extent are the evaluations of CAF consistent between language specialists and content specialists?

RQ2: How do language specialists and content specialists differ in their conceptualization and assessment of CAF in medical writing, and to what extent do their professional perspectives influence the evaluation criteria applied?

RQ3: How do language specialists and content specialists conceptualize CAF in medical writing, and what factors influence their assessments?

RQ4: What types of feedback do language and content specialists provide when evaluating CAF in medical texts?

Methods

This study employs an explanatory sequential mixed methods design, commencing with a quantitative phase. The initial phase aims to assess the consistency of Complexity, Accuracy, and Fluency (CAF) evaluations between language specialists and content specialists (RQ1) and to identify differences in their conceptualizations and assessment criteria influenced by their professional backgrounds (RQ2). During this phase, a structured assessment rubric will be administered to a comprehensive sample of specialists who will evaluate standardized medical writing samples. The resulting quantitative data will be analyzed using statistical techniques, including inter-rater reliability measures and comparative analyses.

Subsequently, a qualitative phase will be conducted to provide a deeper understanding of how language and content specialists conceptualize CAF and the factors that influence their assessments (RQ3), as well as to examine the nature and characteristics of the feedback they provide when evaluating medical texts (RQ4). This phase will involve panel discussions with language and content specialists to interpret the reason of their scoring, priorities and perspectives. The qualitative data will be analyzed thematically to elucidate and contextualize the results.

Adopting an explanatory sequential design facilitates a rigorous and comprehensive investigation, wherein quantitative findings guide the qualitative inquiry, thereby enriching the interpretability and validity of the overall study.⁶⁵

Participants

The study employed purposive sampling to ensure disciplinary representation and assessment literacy diversity. Language specialists were selected based on formal training in applied linguistics and at least five years of EAP instruction experience, while content specialists

had published research or taught clinical writing courses. All participants provided informed consent and were briefed on the CAF dimensions before rating. While purposive sampling limits generalizability, it aligns with exploratory mixed-methods research standards emphasizing depth of expertise over representativeness.

The study employed purposive sampling to ensure disciplinary representation and diversity in assessment literacy. This non-probability approach was selected for its suitability in exploratory mixed-methods designs emphasizing depth of expertise over statistical representativeness. The participants comprised twenty specialists, equally divided between language specialists ($n=10$) and content specialists ($n=10$). The language specialists were instructors formally trained in applied linguistics, possessing expertise in English for Academic Purposes (EAP) and at least five years of experience teaching academic writing. The content specialists were faculty members within clinical and paramedical disciplines who had published research or taught courses in medical English and scientific writing.

All participants were nonnative speakers of English and had taught medical writing or related courses to undergraduate and postgraduate students in medical and health sciences programs. On average, participants had six years of university-level teaching experience, although four members of the cohort had longer records of approximately nine years. The participants were drawn from the medical and paramedical faculties of Tabriz University of Medical Sciences and Tehran University of Medical Sciences, ensuring diverse institutional representation and disciplinary perspectives on Complexity, Accuracy, and Fluency (CAF) in medical writing assessment.

Each participant received an explanation of the study purpose and procedures and gave informed consent prior to participation. They attended a short orientation session on the operationalization of CAF constructs to ensure consistent understanding of key terms across disciplinary backgrounds. This careful selection and pre-briefing process aimed to balance linguistic and disciplinary expertise, supporting credible, interdisciplinary insight into assessor conceptions of CAF in medical writing.

Materials

The assessment of Complexity, Accuracy, and Fluency was conducted by two distinct groups of specialists: language experts and content experts. The analysis was based on a subset of medical texts comprising six short essays, each approximately 200 words in length, authored by advanced second-language (L2) medical students enrolled at Tabriz University of Medical Sciences. These texts were selected from a larger corpus of medical writings produced by a total of forty-four students as part of a writing course. All compositions were completed under controlled, timed conditions within classroom settings, with a duration of

forty minutes allocated for each writing task. To ensure the authenticity and independence of student output, the use of external resources, including dictionaries and reference materials, was strictly prohibited.

The assessment of Complexity, Accuracy, and Fluency (CAF) was carried out by two distinct groups of specialists—language experts and content experts—using a set of medical texts purposefully drawn from a larger corpus to ensure both linguistic comparability and disciplinary authenticity. The primary dataset comprised six short essays (approximately 200 words each) authored by advanced second-language (L2) medical students enrolled at Tabriz University of Medical Sciences. These essays were selected from a larger corpus of forty-four compositions produced within a semester-long medical writing course that required students to compose topic-centered academic texts related to clinical conditions.

The writing tasks were completed under controlled classroom conditions within a forty-minute timeframe, ensuring consistency of context and minimizing extraneous cognitive or linguistic aids. To preserve authenticity and independence of student performance, external reference materials such as dictionaries, internet sources, and discipline-specific manuals were not permitted. The selected sample reflects diversity in textual quality, syntactic structure, and rhetorical cohesion, enabling a balanced representation of proficiency levels necessary for reliable inter-rater evaluation. Moreover, the inclusion of multiple thematic subtopics under the shared clinical theme provides an appropriate balance between content focus and linguistic variation, allowing CAF dimensions to be evaluated across comparable yet distinctive texts.

Although concise by design, the six-text sample aligns with exploratory mixed-methods protocols, emphasizing qualitative interpretative depth over quantitative generalizability. The corpus facilitated triangulated analysis between linguistic features and rater justifications—appropriate for identifying patterns in cross-disciplinary judgment and variation in CAF conceptualization among language and content specialists. This deliberate design ensures methodological transparency and addresses concerns regarding sample narrowness by explicitly linking text selection logic to research intent.

In line with the explanatory sequential mixed-methods design adopted in this study, integration occurred at multiple stages of the research process. First, quantitative findings from the CAF ratings and inter-rater reliability analyses informed the focus of the qualitative phase by identifying dimensions that demonstrated either strong convergence (Accuracy) or variability across rater groups (Complexity and Fluency). These quantitative patterns guided the development of prompts for the panel discussions and the analytic focus of the qualitative coding.

Second, integration was achieved during interpretation, where qualitative thematic findings were used to explain, contextualize, and expand upon quantitative results. Specifically, thematic analysis of raters' written justifications and panel discussions elucidated why Accuracy yielded near-perfect agreement, how Complexity was conceptualized differently depending on disciplinary background, and why Fluency demonstrated greater variability—particularly among content specialists.

Finally, quantitative ICC findings and qualitative themes were brought together through integrative interpretation in the Results and Discussion sections, enabling a more nuanced understanding of cross-disciplinary CAF assessment in medical writing. This approach aligns with best practices in explanatory sequential mixed-methods research, where qualitative data serve an explanatory function rather than merely restating quantitative outcomes.

This selection strategy aligns with exploratory corpus-based designs commonly adopted in CAF research where smaller, contextually controlled samples allow for rich inter-rater and construct-level analysis.^{24,35} Consistent with quality assurance frameworks in medical education⁶⁶, task uniformity, authenticity, and time-bound testing conditions were prioritized to ensure comparability across participants and text types.

CAF Rubric and Rating Procedures

The CAF rubric was systematically developed following analytic rubric design principles that prioritize specific, observable criteria across defined performance levels.¹ Each dimension—Complexity, Accuracy, and Fluency—is operationalized on a six-point scale, with increasing scores reflecting greater linguistic sophistication and disciplinary precision.

Complexity was conceptualized as the degree of syntactic and lexical elaboration within a text, particularly the use of subordination, nominalization, and varied clause structures. Higher levels of performance demonstrated mastery of advanced medical terminology and nuanced sentence constructions, signifying cognitive control over linguistic resources relevant to disciplinary writing. Scores ranged from simple and isolated phrase use (Level 1) to highly controlled, intricate sentence development incorporating precise medical lexis (Level 6).^{25,38}

Accuracy represented the extent to which linguistic forms and medical content adhered to grammatical norms and field-specific correctness. It focuses on grammatical, terminological, and semantic correctness, ranging from frequent errors impeding understanding (level 1) to virtually error-free, fully precise, and reliable content (level 6). This dimension aligns with weighted error scoring models validating communicative effectiveness in academic medical texts.^{2,67}

Fluency describes the global coherence and logical progression of ideas, starting from disjointed, incoherent,

illogical flow (level 1) through to highly coherent, natural flow, professional style (level 6). Concepts of fluency as ease of production and flow in writing are consistent with process-oriented writing measures.²⁷

Each participant's writing sample was rated independently by trained evaluators on the three CAF dimensions using a six-point analytic scale. The descriptors for each level were adapted from evidence-centered assessment principles and contextualized through a pilot study involving both language and special experts. Calibration sessions were conducted prior to data collection to ensure consistency in rater interpretation and scoring. Inter-rater reliability was determined using an intra-class correlation coefficient (ICC) model (two-way random-effects, absolute agreement), yielding coefficients above 0.75, confirming good reliability.

The operationalization of this rubric provides a measurable framework for assessing multidimensional components of writing quality. It establishes construct validity by linking observable linguistic features (e.g., clause diversity, lexical precision, syntactic control) to functional communicative outcomes (e.g., clarity, cohesion, and clinical precision) relevant to professional medical communication. This analytical structure advances methodological transparency and supports reproducibility within interdisciplinary CAF-based research contexts.

Data Collection Procedure

For this study, data were collected during structured one-hour rating sessions held simultaneously for both language specialists and content specialists. At the beginning of session, the study's objectives were clearly explained, and teachers were introduced to the assessment framework designed to evaluate the Complexity, Accuracy, and Fluency of the medical writing samples. After having explained the session's purpose, we presented the rating scale for the assessment of syntactic complexity, a six-point Likert scale derived from the Common European Framework of Reference for Languages (CEFR). This scale was specifically adapted to evaluate complexity, accuracy and fluency in medical writings (see [Appendix 1](#) for the detailed rating scale).

Level 1 of the rating scale is characterized by the presence of simple, isolated phrases and sentences, limited control over a narrow range of basic grammatical structures and sentence patterns, and errors that may impede comprehension. This reflects low complexity, limited accuracy, and disrupted fluency as described within the Complexity, Accuracy, and Fluency (CAF) framework. Conversely, Level 6 denotes the production of a clear, highly accurate, and seamlessly flowing complex text, exhibiting a wide range of cohesive devices and demonstrating proficient control of advanced syntactic forms, indicating high levels of complexity, accuracy, and fluency. Raters were explicitly instructed to focus their

evaluations primarily on syntactic complexity as one dimension of CAF, while excluding assessments related to vocabulary choice, word-level accuracy, and orthographic correctness. To ensure a clear understanding of the scale descriptors and the rating procedure, a representative writing sample was collaboratively reviewed and discussed prior to the commencement of the rating process. Raters were instructed to evaluate each of the three dimensions independently—Complexity, Accuracy, and Fluency—ensuring that judgments on one dimension did not bias ratings on the others. To promote consistency and a common understanding of the criteria and procedures, the panel first reviewed and discussed an example medical writing sample as a group.

Subsequently, each specialist independently assessed six medical texts, each about 200 words long. For each text, raters were asked to:

- Assign separate scores on the six-point scale for Complexity, Accuracy, and Fluency;
- Provide written explanations justifying their scores, referencing specific features of the texts;
- Offer feedback or improvement suggestions targeted at each assessed dimension.

Following the independent evaluations, separate panel discussions were held within each specialist group. During these sessions, raters shared and debated their scoring rationales and feedback strategies for the texts. The discussions were audio-recorded and fully transcribed to enable qualitative analysis of raters' conceptualizations and decision-making processes.

This combined quantitative and qualitative methodology facilitated an in-depth understanding of not only the consistency and variation in rating outcomes but also the professional reasoning behind specialists' assessments of medical writing quality.⁵⁷

Data Analysis

The data analysis consisted of both quantitative and qualitative components to comprehensively examine the specialists' assessments of medical writing.

The statistical procedures used in the data analysis involved both quantitative and qualitative components to comprehensively evaluate specialists' assessments of medical writing.

For the quantitative analysis, intra-class correlation coefficients (ICCs) were calculated separately for the two specialist groups—language specialists and content specialists—based on their ratings using a six-point Likert scale across the three dimensions of Complexity, Accuracy, and Fluency. The ICC is a widely used statistical measure that quantifies the reliability of ratings by assessing the degree of agreement or consistency among raters within the same group. It is computed as the ratio of variance of interest (true variance between subjects) to total variance (including error variance), with values ranging from 0 to 1, where values above 0.8

indicate good to excellent reliability. Two types of ICC models can be applied depending on whether absolute agreement between raters or consistency in ranking order is prioritized. In this study, the ICC provided evidence for the internal consistency and coherence of the rating scales and specialists' judgments in evaluating the medical texts.

Qualitative data consisted of written comments accompanying the ratings and verbatim transcripts of audio-recorded panel discussions with each specialist group. A thematic analysis was conducted on these qualitative materials, systematically coding and categorizing data to identify overarching themes reflecting key evaluative criteria. Major macro-categories extracted included linguistic accuracy, clarity and comprehensibility, use of medical terminology, syntactic complexity, and text organization and coherence. The syntactic complexity category was further subdivided to examine specific linguistic features such as clause complexity, verb phrase construction, phrase order, and noun phrase elaboration, enriching understanding of specialists' perceptions and criteria in medical academic writing evaluation.

Quantitative Analysis

We first evaluated inter-rater reliability within each specialist group—language specialists and content specialists—by calculating intra-class correlation coefficients (ICCs). The ICCs were computed based on the six-point Likert scale ratings that each rater assigned for the three dimensions: Complexity, Accuracy, and Fluency. This statistical measure assessed the degree of consistency and agreement among raters within the same group, providing evidence for the reliability of the rating scales and the coherence of specialists' judgments.

Qualitative Analysis

Subsequently, we systematically analyzed qualitative data, which comprised the specialists' written comments accompanying the rating forms. These comments included both justifications for the numerical scores and constructive feedback aimed at improving the medical texts. Additionally, the audio-recorded panel discussions, conducted separately with each specialist group, were transcribed verbatim to enrich the qualitative dataset.

A comprehensive thematic analysis was carried out on the combined qualitative material. This process involved coding and categorizing data to identify salient themes that reflected the key aspects specialists considered during their evaluations. Emergent macro-categories included:

- Linguistic accuracy
- Clarity and comprehensibility
- Use of medical terminology
- Syntactic complexity
- Text organization and coherence

To gain deeper insight into specialists' perceptions of syntactic complexity, comments within this category

were further disaggregated into subcategories relating to distinct linguistic features such as clause complexity, verb phrase constructions, phrase order, and noun phrase elaboration.

Results

This study analyzed how language specialists and content specialists assessed six medical writing texts on the dimensions of Complexity, Accuracy, and Fluency (CAF) using a six-point Likert scale. Both quantitative reliability and rating data, as well as qualitative thematic data from raters' justifications and group discussions, were examined.

Quantitative Findings

The six medical texts were evaluated using a standardized rubric assessing three dimensions: Complexity, Accuracy, and Fluency. Evaluations were performed independently by a medical specialist and a language specialist to ensure a comprehensive appraisal of both content and linguistic quality.

As shown in Table 1, across the corpus, Accuracy consistently received high ratings (score=5) for Texts 1 through 5 by both evaluators, indicating that these texts provide factually sound, evidence-based, and clinically reliable information. Text 6 was an exception, with scores of 3 from both raters, reflecting several inaccuracies, oversimplifications, and potential misunderstandings, particularly concerning pharmacological information.

Regarding Complexity, scores predominantly fell within

moderate ranges (3 to 4). Texts 1, 3, 4, and 5 demonstrated adequate medical terminology and sentence structures representative of intermediate complexity suitable for medical professionals or educated lay audiences. Text 2 exhibited slightly less complexity, consistent with its lay-oriented explanatory style. Text 6 scored lowest, especially on linguistic complexity (rated 2 by the language specialist), due to simplistic syntax, informal tone, and a lack of technical precision.

In terms of Fluency, the language specialist assigned the highest score (5) to Text 1, recognizing its coherent progression, precise transitions, and polished narrative style. Other texts generally received scores of 4, indicating satisfactory coherence and readability. Text 6 was rated lowest again (2-3), with examiners noting disjointed phrasing, frequent informal expressions, and impaired narrative flow, detracting from professional communication standards.

Taken together, these results indicate that Texts 1 through 5 conform to professional standards for medical writing relevant to general medical audiences or patient education, balancing accessibility with clinical accuracy and linguistic clarity. Text 1 stands out for its superior fluency and balanced complexity, making it a benchmark for effective medical communication. Conversely, Text 6 demonstrates deficiencies across critical evaluation domains, highlighting the need for substantial revision to enhance scientific accuracy and linguistic professionalism.

These findings suggest that while the majority of the texts are appropriate for educational or semi-specialized dissemination about hypertension, additional editorial refinement is warranted for lower-scoring materials to ensure rigorous and precise medical communication.

The inter-rater reliability analysis using Intraclass Correlation Coefficients (ICC) across the Complexity, Accuracy, and Fluency (CAF) dimensions reveals notable differences between language specialists and content specialists in their consistency of ratings.

Table 2 shows that, for the Accuracy dimension, both specialist groups show high levels of agreement, with ICC values of 0.85 for language specialists and 0.91 for content specialists. This suggests that raters from both groups consistently evaluate the accuracy of the texts, indicating a shared understanding and application of accuracy criteria. The slightly higher ICC for content specialists may reflect their focused expertise in the content-related correctness of the materials being assessed.

In the Fluency dimension, the ICC is 0.82 for language specialists but drops to 0.63 for content specialists. This indicates that language specialists maintain a higher level

Table 1. Mean Likert Scores and Standard Deviations for Complexity, Accuracy, and Fluency by Specialist Group across Six Medical Texts

Text	Evaluator	Complexity (C)	Accuracy (A)	Fluency (F)
1	Medical Specialist	4	5	4
	Language Specialist	4	5	5
2	Medical Specialist	3	5	4
	Language Specialist	3	5	4
3	Medical Specialist	4	5	4
	Language Specialist	4	5	4
4	Medical Specialist	4	5	4
	Language Specialist	4	5	4
5	Medical Specialist	4	5	4
	Language Specialist	4	5	4
6	Medical Specialist	3	3	3
	Language Specialist	2	3	2

Note. Ratings are on a six-point Likert scale (1=lowest, 6=highest performance level).

Table 2. Inter-Rater Reliability (ICC) by Specialist Group and CAF Dimension

Dimension	Language Specialists ICC	Content Specialists ICC	Interpretation
Accuracy	0.85	0.91	High agreement in both groups
Fluency	0.82	0.63	Higher consistency among language specialists
Complexity	0.74	0.68	Moderate agreement; somewhat lower for content specialists

of consistency when assessing fluency aspects such as the flow, readability, and naturalness of language. The lower ICC for content specialists suggests greater variability or less consensus in their evaluation of fluency, possibly because fluency is less central to their content-focused expertise.

Regarding Complexity, both groups exhibit moderate agreement, with language specialists at 0.74 and content specialists at 0.68. These moderate ICC values imply a reasonable but less robust consensus compared to Accuracy. The slightly higher agreement among language specialists might be due to their more refined sensitivity to linguistic intricacies and syntactic or lexical complexity. Content specialists might show somewhat lower consistency here, perhaps because linguistic complexity is not their primary evaluative focus.

Overall, these results highlight that while both groups are reliable raters for Accuracy, language specialists demonstrate stronger reliability in Fluency and Complexity assessments. This pattern suggests the importance of involving language specialists when evaluating nuanced linguistic features such as fluency and complexity, whereas content specialists are equally reliable for accuracy judgments related to content fidelity.

Qualitative Findings

Thematic analysis of raters' written justifications and panel discussions was conducted to explore in depth the evaluators' perspectives on the three key dimensions of the rubric: Complexity, Accuracy, and Fluency (CAF). This qualitative inquiry complemented the quantitative scores by elucidating underlying rationales, interpretive nuances, and points of consensus or discrepancy, as summarized in Table 3.

The analysis followed a rigorous, systematic approach consistent with established qualitative research guidelines.⁶⁸ Iterative coding was performed on textual data from individual rater comments and group discussions, allowing themes to emerge inductively while remaining anchored to the CAF framework. The process included familiarization with the data, open coding, theme development, and refinement through collaborative researcher reflection, ensuring the reliability

and depth of thematic insights.

Accuracy emerged as the dimension with the most consistent agreement and was unanimously regarded as the foundational criterion for evaluating medical texts. Both medical and linguistic evaluators emphasized the necessity of factual correctness, clinical reliability, and precise terminology to uphold informational integrity. For example, one rater noted, "The text correctly presents the core clinical definitions and treatment guidelines without any notable errors, supporting its reliability for educational use."

Complexity was characterized by convergence on the importance of appropriately used medical terminology and moderately complex syntactic constructions. However, raters diverged regarding the optimal level of complexity for their intended readership. The medical specialist prioritized clinical depth and pathophysiological detail, whereas the language specialist highlighted lexical variety and syntactic sophistication. A language specialist commented, "Certain sections lacked deeper explanation or clinical nuance that would have elevated the content for specialist readers," underscoring the nuanced nature of this dimension.

The Fluency dimension elicited the greatest variability in perspectives. While all raters acknowledged the importance of clarity, coherence, and logical structure, the linguistic expert placed greater emphasis on stylistic polish, narrative flow, and transitional devices. Conversely, the medical specialist focused primarily on clarity that facilitates clinical reasoning and information transfer. One evaluator remarked, "The flow of information is clear, with logical progression and well-connected sentences, facilitating reader comprehension," while another observed in a less fluent passage, "The text contains repetitive phrasing and informal language that interrupts the smooth reading experience and reduces professional tone."

A recurrent theme across discussions was the challenge of balancing clinical precision with accessibility. Evaluators highlighted the tension between maintaining technical accuracy and crafting texts that remain comprehensible and engaging for patients or lay readers. This trade-off was succinctly captured by a rater who stated, "Balancing

Table 3. Summary of Thematic Findings on CAF Dimensions

CAF Dimension	Converging Themes	Diverging Themes	Representative Quotes
Accuracy	- Primacy of factual correctness and clinical reliability	- Minimal divergence; largely objective and agreed upon	"The text correctly presents the core clinical definitions and treatment guidelines without notable errors."
Complexity	- Importance of relevant medical terminology and syntactic variety	- Optimal complexity level varies depending on intended audience	"Certain sections lacked deeper explanation or clinical nuance that would have elevated the content..."
Fluency	- Necessity of clear, coherent, and logically structured prose	- Differing emphasis on stylistic polish and narrative flow	"The flow of information is clear, with logical progression..."; "Repetitive phrasing interrupts smooth reading."
Technical vs Accessibility	- Challenge in balancing clinical precision with readability	- Trade-off between maintaining accuracy and simplifying language	"Balancing the need for technical accuracy with readability is a persistent challenge."
Evaluator Perspectives	- Recognition of influence of disciplinary background	- Clinical depth prioritized by medical specialist; linguistic style by language specialist	"My background as a clinician leads me to prioritize clinical depth and detail." "As a language specialist, I focus more on syntactic variety and narrative coherence."

the need for technical accuracy with readability is a persistent challenge, especially when targeting patient education materials.”

Finally, the analysis illuminated how disciplinary backgrounds influence evaluative judgments. The medical specialist’s clinical experience shaped prioritization of content depth and medical correctness, while the language specialist’s expertise oriented them toward linguistic clarity and stylistic consistency. As one participant reflected, “My background as a clinician leads me to prioritize clinical depth and detail,” compared to another’s comment, “As a language specialist, I focus more on syntactic variety and narrative coherence, which can result in divergent judgments on complexity and fluency.”

Therefore, the thematic analysis reveals the multidimensional and multifaceted nature of medical text evaluation. It underscores the complementary role of combined medical and linguistic expertise in achieving comprehensive quality assessments. These qualitative insights enrich the interpretation of the quantitative inter-rater reliability results and highlight the nuanced balancing acts inherent in effective medical communication.

When interpreted together, the quantitative and qualitative findings reveal a coherent explanatory pattern. The near-perfect ICC values for Accuracy were explained qualitatively by raters’ shared belief that factual correctness and clinical reliability constitute non-negotiable criteria in medical writing. In contrast, the moderate ICC values observed for Complexity and Fluency were illuminated by qualitative evidence showing that these constructs were interpreted through distinct professional lenses. Language specialists associated Complexity with syntactic variety and lexical sophistication, whereas content specialists linked it to conceptual depth and clinical nuance. Similarly, Fluency was consistently associated with stylistic coherence by language specialists but viewed primarily as functional clarity by content specialists. Thus, the qualitative findings do not merely replicate quantitative trends but explicate the disciplinary reasoning that produced them, demonstrating the explanatory value of the mixed-methods design.

Discussion

This study explored how language specialists and content specialists conceptualize and assess CAF in medical writing, bringing new insights into the convergences and divergences in their evaluations. The employment of an explanatory sequential mixed methods design allowed both quantitative reliability measurement and qualitative thematic analysis to reveal the multifaceted nature of CAF assessment from professional perspectives.

The findings of this study should be interpreted in light of the intentionally limited corpus size. The quantitative component was based on six medical texts, a scope that is appropriate for exploratory mixed-methods

research aimed at examining rater behavior, construct interpretation, and interdisciplinary assessment practices rather than statistical generalization. Accordingly, the quantitative analyses are best understood as illustrative rather than confirmatory.

While the small corpus restricts inferential claims about broader populations of medical texts or raters, it enabled fine-grained analysis of evaluative consistency and facilitated deep qualitative exploration of assessor reasoning. This trade-off aligns with established mixed-methods and CAF research traditions, which emphasize interpretive depth and construct validity over representativeness in early-stage or theory-building studies.

Future research should expand the corpus size and include a wider range of genres, proficiency levels, and medical specialties to test the stability of the observed patterns. Nevertheless, the present findings provide methodologically transparent, theoretically grounded insights into how language and content specialists conceptualize and apply CAF in medical writing assessment.

Consistent with prior research, high inter-rater reliability was observed for the Accuracy dimension across both specialist groups, with content specialists exhibiting slightly higher agreement (ICC=0.91). This aligns with established findings positioning accuracy as the foundational pillar in medical text evaluation due to its crucial role in maintaining clinical correctness and factual integrity.⁶⁹ Both groups demonstrated a shared understanding that precise, evidence-based content is non-negotiable in medical communications, confirming the importance of this construct in medical writing assessment frameworks. However, reliability for the Complexity and Fluency dimensions showed moderate agreement, with language specialists displaying greater consistency, reflecting their expertise in linguistic features such as syntactic variety, cohesion, and narrative flow. This echoes findings by Kuiken and Vedder,⁵⁷ who reported that professional background significantly influences rater attention to particular construct dimensions in second language writing. The moderate ICCs on CAF Complexity also resonate with literature revealing difficulties in reliably assessing syntactic and phrasal elaboration, especially among content-oriented raters.⁷⁰

Qualitative data expanded understanding by highlighting how content specialists foregrounded the clinical relevance and terminological precision while language specialists foregrounded linguistic intricacies such as syntactic complexity and textual coherence. These complementary emphases confirm Biber, Gray²² affirmation of the critical role that advanced syntactic structures and lexical variety play in academic medical prose. Similarly, Pallotti’s²⁴ research on second language acquisition supports the notion that differential

conceptualizations of CAF are grounded in disciplinary paradigms and pedagogical objectives.

Both groups tended to center feedback on deficits and errors, a characteristic observed in multiple studies on medical writing pedagogy and SLA.^{24,69} The preferential focus on accuracy by content specialists concurs with Gattrell et al,⁴⁹ who demonstrated that clinical expertise underpins reliable medical content evaluation. Language specialists' prominence in fluency and syntactic complexity rating echoes broader language assessment literature underscoring their roles as custodians of textual coherence and stylistic sophistication.²²

Our finding that phrasal complexity and coordination received limited assessment attention replicates concerns raised by Norris and Ortega⁷ and others, who advocate for enhanced training aimed at recognizing and evaluating these advanced linguistic features essential in professional writing. The lack of focus on these complexity subdomains may negatively impact the completeness of evaluations and subsequent writing feedback.

The tension between maintaining clinical precision and ensuring reader accessibility is a dominant theme across medical communication research.¹⁸ This tension demands sophisticated assessor judgment and balanced textual strategies to meet the needs of both expert and lay audiences, an issue underscored in our study. The demonstrated differences in CAF conceptualization suggest that interdisciplinary panels of language and content specialists represent a best practice for medical writing assessment. Such panels can collectively monitor content accuracy, linguistic complexity, and fluency, thereby producing robust, multidimensional evaluations that individual raters might miss. Training specialists to recognize and evaluate syntactic and phrasal complexity with greater sensitivity is critical for advancing assessment precision and feedback efficacy. This need aligns with the calls from Biber, Gray²² and Norris & Ortega⁷ regarding specialized training focused on linguistic intricacies crucial to academic and professional medical writing.

Furthermore, fostering a positive feedback culture that balances error correction with recognition of developmental gains can enhance motivation and accelerate writing proficiency, echoing Pallotti's²⁴ SLA recommendations and clinical education best practices.⁶⁹ Lastly, the ongoing challenge of balancing technical accuracy with readability suggests that medical writing curricula and assessment frameworks should explicitly address these dual demands. This would support the production of texts that simultaneously satisfy professional content standards and accessibility requirements for diverse audiences.¹⁸

Recent developments in AI-driven medical text evaluation provide a complementary lens to interpret our findings. Automated approaches, such as LLM-as-a-Judge systems⁹, show promising alignment with human judgments on factual accuracy but remain

challenged by reproducibility and standardization issues. Systematic analyses of AI-based medical text evaluation¹⁰ emphasize variability in evaluation criteria and the need for transparent operationalization of assessment rubrics. Parallel research in medical education and generative AI applications^{11,13} highlights the importance of clearly defined, replicable methodologies. Our CAF-based framework parallels these objectives by offering a structured, replicable approach to assess Accuracy, Complexity, and Fluency, confirming high reliability for Accuracy while highlighting the nuanced judgments required for Complexity and Fluency. These findings suggest that interdisciplinary human evaluation can both inform and validate AI-driven methods, ensuring that clinical precision and linguistic sophistication are appropriately weighted in automated assessment systems.

Conclusion

This study evaluated medical texts on hypertension using expert ratings of Complexity, Accuracy, and Fluency (CAF). Accuracy emerged as the most critical and consistently rated dimension, underscoring its foundational role in medical communication. Complexity and fluency showed more variation, reflecting differing disciplinary perspectives and subjective judgments.

The combination of quantitative and qualitative findings highlights the importance of interdisciplinary evaluation, integrating clinical expertise and linguistic insight. Balancing clinical precision with accessibility remains a key challenge in medical text development.

The quantitative results should be viewed as exploratory indicators of rater agreement that gain explanatory power through integration with qualitative findings, rather than as evidence for broad statistical generalization.

Future work should expand this framework to diverse texts and audiences, potentially incorporating AI tools to complement human judgment. Ultimately, this approach supports the creation of clear, accurate, and reader-friendly health education materials to improve patient understanding and outcomes.

Acknowledgments

The authors would like to express their sincere gratitude to the language specialists and content experts from Tabriz University of Medical Sciences and Tehran University of Medical Sciences who generously participated in the rating sessions and panel discussions, providing invaluable insights that formed the core of this study. We also extend our thanks to the advanced L2 medical students at Tabriz University of Medical Sciences whose writing samples were essential for this research. Special appreciation goes to our colleagues at the Faculty of Foreign Languages and Literatures, University of Tehran, for their constructive feedback and support during the development of the CAF rubric and data analysis.

Authors' Contribution

Conceptualization: Sayyed Mohammad Alavi, Neda Rahmatvand.
Data Curation: Neda Rahmatvand.
Investigation: Sayyed Mohammad Alavi, Neda Rahmatvand.
Methodology: Sayyed Mohammad Alavi.

Project Administration: Sayyed Mohammad Alavi.
 Resources: Neda Rahmatvand.
 Software: Neda Rahmatvand.
 Supervision: Sayyed Mohammad Alavi.
 Writing—Original Draft: Sayyed Mohammad Alavi, Neda Rahmatvand.
 Writing—Review & Editing: Sayyed Mohammad Alavi.

Competing Interests

The authors declare that they have no competing interests.

Ethical Approval

All participants provided informed consent prior to their engagement in the study. They were informed of the voluntary nature of their involvement and assured that they could withdraw at any time without facing any repercussions. Participant confidentiality was upheld throughout the process, with all data anonymized to ensure privacy protection.

Funding

Not applicable.

References

- Barrot JS, Agdeppa JY. Complexity, accuracy, and fluency as indices of college-level L2 writers' proficiency. *Assess Writ* 2021;47:100510. doi:[10.1016/j.asw.2020.100510](https://doi.org/10.1016/j.asw.2020.100510)
- Bulté B, Housen A. Defining and operationalising L2 complexity. In: *Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA*. John Benjamins; 2012. p. 21-46.
- Wigglesworth G, Storch N. Pair versus individual writing: effects on fluency, complexity and accuracy. *Lang Test* 2009;26(3):445-66. doi:[10.1177/0265532209104670](https://doi.org/10.1177/0265532209104670)
- Zhang S, Zhang H, Zhang C. A Dynamic systems study on complexity, accuracy, and fluency in English writing development by Chinese university students. *Front Psychol* 2022;13:787710. doi:[10.3389/fpsyg.2022.787710](https://doi.org/10.3389/fpsyg.2022.787710)
- Fazilatfar A, Kasiri F, Nowbakht M. The Comparative effects of planning time and task conditions on the complexity, accuracy and fluency of L2 writing by EFL learners. *Iran J Lang Teach Res* 2020;8(1):93-110. doi:[10.30466/ijltr.2020.120809](https://doi.org/10.30466/ijltr.2020.120809)
- Phuoc VD, Barrot JS. Complexity, accuracy, and fluency in L2 writing across proficiency levels: a matter of L1 background? *Assess Writ* 2022;54:100673. doi:[10.1016/j.asw.2022.100673](https://doi.org/10.1016/j.asw.2022.100673)
- Norris JM, Ortega L. Towards an organic approach to investigating CAF in instructed SLA: the case of complexity. *Appl Linguist* 2009;30(4):555-78. doi:[10.1093/applin/amp044](https://doi.org/10.1093/applin/amp044)
- Vercellotti ML. The development of complexity, accuracy, and fluency in second language performance: a longitudinal study. *Appl Linguist* 2017;38(1):90-111. doi:[10.1093/applin/amv002](https://doi.org/10.1093/applin/amv002)
- Croxford E, Gao Y, First E, Pellegrino N, Schnier M, Caskey J, et al. Automating evaluation of AI text generation in healthcare with a large language model (LLM)-as-a-judge. *medRxiv* [Preprint]. May 6, 2025. Available from: <https://www.medrxiv.org/content/10.1101/2025.04.22.25326219v2>.
- Alzarea AI, Ishaqui AA, Maqsood MB, Alanazi AS, Alsaidan AA, Mallhi TH, et al. Evaluating AI performance in infectious disease education: a comparative analysis of ChatGPT, Google Bard, Perplexity AI, Microsoft Copilot, and Meta AI. *Front Med (Lausanne)* 2025;12:1679153. doi:[10.3389/fmed.2025.1679153](https://doi.org/10.3389/fmed.2025.1679153)
- Crossnohere NL, Elsaid M, Paskett J, Bose-Brill S, Bridges JFP. Guidelines for artificial intelligence in medicine: literature review and content analysis of frameworks. *J Med Internet Res* 2022;24(8):e36823. doi:[10.2196/36823](https://doi.org/10.2196/36823)
- Roveta A, Castello LM, Massarino C, Francese A, Ugo F, Maconi A. Artificial intelligence in medical education: a narrative review on implementation, evaluation, and methodological challenges. *AI* 2025;6(9):227. doi:[10.3390/ai6090227](https://doi.org/10.3390/ai6090227)
- Akyon SH, Akyon FC, Camyar AS, Hızlı F, Sari T, Hızlı Ş. Evaluating the capabilities of generative AI tools in understanding medical papers: qualitative study. *JMIR Med Inform* 2024;12:e59258. doi:[10.2196/59258](https://doi.org/10.2196/59258)
- Block D. Revisiting the gap between SLA researchers and language teachers. *Links Lett* 2000(7):129-43.
- Behrahi L, Khalaji HR, Rajabi P. A mixed methods study of the impact of online teacher mediation and learner scaffolding via Google Docs on EFL learners' writing complexity, accuracy, and fluency. *Technol Assist Lang Educ* 2024;2(3):1-24. doi:[10.22126/tale.2024.10829.1050](https://doi.org/10.22126/tale.2024.10829.1050)
- Kim S, Yang JW, Lim J, Lee S, Ihm J, Park J. The impact of writing on academic performance for medical students. *BMC Med Educ* 2021;21(1):61. doi:[10.1186/s12909-021-02485-2](https://doi.org/10.1186/s12909-021-02485-2)
- Dergaa I, Ben Saad H, Glenn JM, Ben Aissa M, Taheri M, Swed S, et al. A thorough examination of ChatGPT-3.5 potential applications in medical writing: a preliminary study. *Medicine (Baltimore)* 2024;103(40):e39757. doi:[10.1097/md.00000000000039757](https://doi.org/10.1097/md.00000000000039757)
- Khojasteh L, Kafipour R, Pakdel F, Mukundan J. Empowering medical students with AI writing co-pilots: design and validation of AI self-assessment toolkit. *BMC Med Educ* 2025;25(1):159. doi:[10.1186/s12909-025-06753-3](https://doi.org/10.1186/s12909-025-06753-3)
- Casanave CP. *Controversies in Second Language Writing: Dilemmas and Decisions in Research and Instruction*. University of Michigan Press; 2013.
- Larsen-Freeman D. An ESL index of development. *TESOL Q* 1978;12(4):439-48. doi:[10.2307/3586142](https://doi.org/10.2307/3586142)
- Housen A, Kuiken F. Complexity, accuracy, and fluency in second language acquisition. *Appl Linguist* 2009;30(4):461-73. doi:[10.1093/applin/amp048](https://doi.org/10.1093/applin/amp048)
- Biber D, Gray B, Poonpon K. Should we use characteristics of conversation to measure grammatical complexity in L2 writing development?. *Tesol Quarterly* 2011;45(1):5-35. doi:<https://doi.org/10.5054/tq.2011.244483>
- Kyle K, Crossley SA. Measuring syntactic complexity in L2 writing using fine-grained clausal and phrasal indices. *The Modern Language Journal* 2018;102(2):333-49.
- Pallotti G. CAF: defining, refining and differentiating constructs. *Appl Linguist* 2009;30(4):590-601. doi:[10.1093/applin/amp045](https://doi.org/10.1093/applin/amp045)
- Lu X. A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Q* 2011;45(1):36-62. doi:[10.5054/tq.2011.240859](https://doi.org/10.5054/tq.2011.240859)
- Taguchi E, Takayasu-Maass M, Gorsuch GJ. Developing reading fluency in EFL: How assisted repeated reading and extensive reading affect fluency development. *Reading in a foreign language* 2004;16(2):70-96.
- Plakans L, Gebril A. A close investigation into source use in integrated second language writing tasks. *Assessing Writing*. 2012;17(1):18-34.
- Larsen-Freeman D. The emergence of complexity, fluency, and accuracy in the oral and written production of five Chinese learners of English. *Appl Linguist* 2006;27(4):590-619. doi:[10.1093/applin/aml029](https://doi.org/10.1093/applin/aml029)
- Skehan P. *A Cognitive Approach to Language Learning*. Oxford University Press; 1998.
- Parkinson J, Musgrave J. Development of noun phrase complexity in the writing of English for Academic Purposes students. *J Engl Acad Purp* 2014;14:48-59. doi:[10.1016/j.jeap.2013.12.001](https://doi.org/10.1016/j.jeap.2013.12.001)

31. Crossley SA, McNamara DS. Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners. *J Second Lang Writ* 2014;26:66-79. doi:10.1016/j.jslw.2014.09.006
32. Coxhead A. A new academic word list. *TESOL Q* 2000;34(2):213-38. doi:10.2307/3587951
33. Douglas SR. The lexical breadth of undergraduate novice level writing competency. *Can J Appl Linguist* 2013;16(1):152-70.
34. Kong N. Developing Meaning-Based Complexity Indices Based on Cognitive-Functional Connectives and Validating Their Predictive Utility in Automated Writing Evaluation. *Rers Sq [Preprint]*. September 1, 2025. Available from: <https://www.researchsquare.com/article/rs-7247858/v1>.
35. Ortega L. Syntactic complexity measures and their relationship to L2 proficiency: a research synthesis of college-level L2 writing. *Appl Linguist* 2003;24(4):492-518. doi:10.1093/applin/24.4.492
36. Khushik GA, Huhta A. Syntactic complexity in Finnish-background EFL learners' writing at CEFR levels A1-B2. *European Journal of Applied Linguistics* 2022;10(1): 142-184. doi:10.1515/eujal-2021-0011
37. Lu X. Automatic analysis of syntactic complexity in second language writing. *Int J Corpus Linguist* 2010;15(4):474-96. doi:10.1075/ijcl.15.4.02lu
38. Zhang X, Lu X. Testing the relationship of linguistic complexity to second language learners' comparative judgment on text difficulty. *Lang Learn* 2024;74(3):672-706. doi:10.1111/lang.12633
39. Wolfe-Quintero KE, Inagaki S, Kim HY. *Second Language Development in Writing: Measures of Fluency, Accuracy, & Complexity*. University of Hawaii Press; 1998.
40. Neumann H. Teacher assessment of grammatical ability in second language academic writing: a case study. *J Second Lang Writ* 2014;24:83-107. doi:10.1016/j.jslw.2014.04.002
41. Cumming A, Kantor R, Baba K, Eouanzoui K, Erdosy U, Jamse M. Analysis of discourse features and verification of scoring levels for independent and integrated prototype written tasks for the new TOEFL®. *ETS Res Rep Ser* 2005;2005(1):i-77. doi:10.1002/j.2333-8504.2005.tb01990.x
42. Gebriel A, Plakans L. Toward a transparent construct of reading-to-write tasks: the interface between discourse features and proficiency. *Lang Assess Q* 2013;10(1):9-27. doi:10.1080/15434303.2011.642040
43. Foster P, Wigglesworth G. Capturing accuracy in second language performance: the case for a weighted clause ratio. *Annu Rev Appl Linguist* 2016;36:98-116. doi:10.1017/s0267190515000082
44. van Mens HJT, Duijm RD, Nienhuis R, de Keizer NF, Cornet R. Towards an adoption framework for patient access to electronic health records: systematic literature mapping study. *JMIR Med Inform* 2020;8(3):e15150. doi:10.2196/15150
45. Ivanov S, Johansson A, Waldmann C. Writing fluency and digital source use in authentic academic writing. *HumaNetten* 2022(48):262-89. doi:10.15626/hn.20224811
46. Johnson MD. Cognitive task complexity and L2 written syntactic complexity, accuracy, lexical complexity, and fluency: a research synthesis and meta-analysis. *J Second Lang Writ* 2017;37:13-38. doi:10.1016/j.jslw.2017.06.001
47. Yang MC, Kim MJ, Lee HG, Rim HC. Assessing Writing Fluency of Non-English-Speaking Student for Automated Essay Scoring-How to Automatically Evaluate the Fluency in English Essay. In: *Proceedings of the 4th International Conference on Computer Supported Education (CSEDU-2012)*. SCITEPRESS; 2012. p. 83-7.
48. El Burai Felix S, Yusuf H, Ritchey M, Romano S, Namulanda G, Wilkins N, et al. A standard framework for evaluating large health care data and related resources. *MMWR Suppl* 2024;73(3):1-13. doi:10.15585/mmwr.su7303a1
49. Gattrell WT, Hopewell S, Young K, Farrow P, White R, Wager E, et al. Professional medical writing support and the quality of randomised controlled trial reporting: a cross-sectional study. *BMJ Open* 2016;6(2):e010329. doi:10.1136/bmjopen-2015-010329
50. Mazgutova D, Kormos J. Syntactic and lexical development in an intensive English for Academic Purposes programme. *J Second Lang Writ* 2015;29:3-15. doi:https://doi.org/10.1016/j.jslw.2015.06.004
51. Tavakoli P, Foster P. Task design and second language performance: the effect of narrative type on learner output. *Lang Learn* 2011;61(s1):37-72. doi:10.1111/j.1467-9922.2011.00642.x
52. Yan Z, Li Z, Panadero E, Yang M, Yang L, Lao H. A systematic review on factors influencing teachers' intentions and implementations regarding formative assessment. *Assess Educ* 2021;28(3):228-60. doi:10.1080/0969594x.2021.1884042
53. Barnes N, Fives H, Dacey CM. US teachers' conceptions of the purposes of assessment. *Teach Teach Educ* 2017;65:107-16. doi:10.1016/j.tate.2017.02.017
54. Guo Q. Exploring teacher assessment conceptions in University EFL writing instruction. *Sage Open* 2025;15(1):21582440251324746. doi:10.1177/21582440251324746
55. Abd Halim H, Hamzah MI, Zulkifli H. Secondary school teachers' conceptions of assessment: a systematic review. *Int J Educ Res* 2024;124:102311. doi:10.1016/j.ijer.2023.102311
56. Zheng Y, Xu J. Unpacking the impact of teacher assessment approaches on student writing engagement: a survey of university learners across different languages. *Assess Eval High Educ* 2023;48(8):1240-53. doi:10.1080/02602938.2023.2219431
57. Kuiken F, Vedder I. Investigating teachers' perceptions of syntactic complexity in L2 academic writing. *Instructed Second Language Acquisition* 2019;3(2):228-48. doi:10.1558/isla.37977
58. Kuhi D, Rasuli MA, Deylami Z. The effect of type of writing on accuracy, fluency and complexity across proficiency. *Procedia Soc Behav Sci* 2014;98:1036-45. doi:10.1016/j.sbspro.2014.03.514
59. Wang G, Wang H, Sun X, Wang N, Wang L. Linguistic complexity in scientific writing: a large-scale diachronic study from 1821 to 1920. *Scientometrics* 2023;128(1):441-60. doi:10.1007/s11192-022-04550-z
60. Sharma S. How to become a competent medical writer? *Perspect Clin Res* 2010;1(1):33-7.
61. Weigle SC. *Assessing Writing*. Cambridge University Press; 2002.
62. Han Q. *Rater Cognition in L2 Speaking Assessment: A Review of the Literature*. 2015. Available from: <https://academiccommons.columbia.edu/doi/10.7916/D82R53MF>.
63. Eldin A, Ahmed M. Bridging language barriers: addressing complexity, accuracy, and fluency challenges of Egyptian researchers in writing research for international publication. *Transcult J Humanit Soc Sci* 2025;6(3):107-26. doi:10.21608/tjhss.2025.376419.1319
64. Evuarherhe O, Gattrell W, White R, Winchester CC. Professional medical writing support and the quality, ethics and timeliness of clinical trial reporting: a systematic review. *Res Integr Peer Rev* 2019;4(1):14. doi:10.1186/s41073-019-0073-7
65. Ary D, Jacobs LC, Razavieh A. *Introduction to Research in Education*. 9th ed. Cengage Learning; 1972.
66. Karthikeyan S, O'Connor E, Hu W. Motivations of assessment item writers in medical programs: a qualitative study. *BMC Med Educ* 2020;20(1):334. doi:10.1186/s12909-020-02229-8
67. Barroga E, Mitoma H. Improving scientific writing skills and

publishing capacity by developing university-based editing system and writing programs. *J Korean Med Sci* 2019;34(1):e9. doi:10.3346/jkms.2019.34.e9

68. Braun V, Clarke V. Using thematic analysis in psychology. *Qual Res Psychol* 2006;3(2):77-101.

69. Astaneh B, Raeisi Shahraki H, Astaneh V, Guyatt G. Assessment of confidence in medical writing: Development and validation of the first trustworthy measurement tool. *PLoS One* 2024;19(4):e0302299. doi:10.1371/journal.pone.0302299

70. Révész A, Gurzynski-Weiss L. Teachers’ perspectives on second language task difficulty: insights from think-alouds and eye tracking. *Annu Rev Appl Linguist* 2016;36:182-204. doi:10.1017/S0267190515000124

Appendix 1. Rating scale

Score	Complexity	Accuracy	Fluency
1	Simple, isolated phrases, poor use of medical vocabulary	Frequent grammatical and terminology errors, impeding understanding	Disjointed, incoherent, illogical flow
2	Basic sentence linking, limited varied structures	Some errors but meaning often recoverable	Limited coherence, awkward transitions
3	Some subordination or noun phrase elaboration; basic medical terms	Occasional errors but reader can infer meaning	Reasonably clear flow with minor lapses
4	Clear, controlled complex sentence usage and terminology	Minor errors that do not impede comprehension	Smooth text, logical sequence
5	Well-structured complex sentences, appropriate and varied medical vocabulary	Accurate grammar and content, precise terminology	Fluent, effective use of cohesion
6	Highly sophisticated complex constructions and advanced medical terms	Virtually error-free, fully precise and reliable content	Highly coherent, natural flow, professional style