GAZİANTEP UNIVERSITY
JOURNAL OF SOCIAL SCIENCES

Journal homepage: http://dergipark.org.tr/tr/pub/jss

Derleme Makalesi ● Review Article

Linguistic Dimensions of L2 Performance: Complexity, Accuracy and Fluency

İkinci Dil Performansının Dilbilimsel Boyutları: Karmaşıklık, Hatasızlık ve Akıcılık

Aysel ŞAHİN KIZIL[a]*

[a] Assoc. Prof., İzmir Bakırçay University, School of Foreign Langauges, İzmir / TÜRKİYE
ORCID: 0000-0001-6277-6208

MAKALE BİLGİSİ

Makale Geçmişi:
Başvuru tarihi: 12 Ağustos 2024
Kabul tarihi: 30 Ağustos 2024

Anahtar Kelimeler:
Karmaşıklık,
Hatasızlık,
Akıcılık,
İkinci dil.

ÖZ

Karmaşıklık, Hatasızlık ve Akıcılık (KHA) modeli, ikinci dil edinimi (SLA) araştırmalarındaki temel modeller arasında yer almakta ve dil performansının çok yönlü doğasını analiz etmek için sağlam bir çerçeve sunmaktadır. Bu çalışma; karmaşıklık, hatasızlık ve akıcılık modelinin, ikinci dil öğrenenlerin dil gelişimini değerlendirmedeki kritik rolünü vurgulayarak modelin kapsamlı bir incelemesini sunmayı amaçlamaktadır. Makale, modeldeki dilbilimsel karmaşıklık, hatasızlık ve akıcılık bileşenlerini süreç içerisinde gelişimlerine kısaca değinerek her bir bileşenin ayrıntılı bir tanımını sunmaktadır. İlgili alan yazında rapor edilen çeşitli deneysel çalışmaların ve metodolojik yaklaşımların gözden geçirildiği bu makalede, ayrıca, dil performansının dilbilimsel yönlerinin ikinci dil araştırmalarında nicel ve nitel yöntemlerle nasıl değerlendirildiğine dair bir analiz sunulmaktadır. Çalışmada karmaşıklık, hatasızlık ve akıcılığı ölçmek için kullanılan farklı metriklerin etkililiği eleştirel bir şekilde değerlendirilmektedir. Ayrıca, bu ölçümlerin ikinci dil yeterliliğini ve dil üretiminin altında yatan bilişsel süreçleri anlamaya yönelik daha geniş etkileri araştırılmaktadır. Son olarak, çalışma mevcut literatürdeki boşlukları belirlemekte ve karmaşıklık, hatasızlık ve akıcılık ölçümlerinin güvenilirliğini ve geçerliliğini artırmaya yönelik gelecekteki araştırma yönlerini özetlemektedir. KHA modelinin hem teorik hem de pratik uygulamalarını derinlemesine inceleyerek, bu makale ikinci dil performans değerlendirmesine dair daha derin bir içgörüye katkıda bulunmaktadır. Bu boyutların ölçülmesi için daha rafine ve kapsamlı yaklaşımlar geliştirilmesinin önemini vurgulayarak, ikinci dil araştırmalarının ilerlemesine ve dil öğretim metodolojilerinin etkinliğini artırmaya yönelik katkı sunmaktadır.

ARTICLE INFO

Article History:
Received: August 12, 2024
Accepted: August 30, 2024

Keywords:
Complexity,
Accuracy,
Fluency,
Second language.

ABSTRACT

The Complexity, Accuracy, and Fluency (CAF) framework is among the fundamental models in second language acquisition (SLA) research, providing a robust structure for analysing the multifaceted nature of language performance. This paper sets out to offer a comprehensive examination of the CAF model, emphasizing its critical role in evaluating the linguistic development of language learners. It thoroughly defines each component—complexity, accuracy, and fluency—while highlighting the intricate challenges involved in measuring these dimensions. Through a review of various empirical studies and methodological approaches, the paper presents a nuanced analysis of how these aspects of language performance are quantified and assessed within SLA research. This study critically evaluates the effectiveness of different metrics used to measure complexity, accuracy, and fluency. It also explores the broader implications of these measurements for understanding L2 proficiency and the cognitive processes underlying language production. Furthermore, the study identifies existing gaps in the literature and proposes future research directions aimed at enhancing the reliability and validity of CAF measurements. By providing an in-depth analysis of both the theoretical and practical applications of the CAF model, this paper contributes to a deeper understanding of L2 performance assessment. It underlines the importance of developing more refined and comprehensive approaches to measure these dimensions, with the ultimate goal of advancing SLA research and improving the effectiveness of language teaching methodologies.

* *Sorumlu yazar/Corresponding author*.
e-posta: aysel.sahinkizil@bakircay.edu.tr

Introduction

The acquisition of a second language (L2) is regarded a multifaceted, complex process characterized by the gradual development of multiple linguistic competencies. This intricate process involves mastering a number of linguistic components including phonological, morphological, syntactic, lexical, and phraseological features of the language being learned (Housen & Kuiken, 2009). Because language is inherently complex, learner progress cannot be fully captured by evaluating performance in any single subsystem alone (Larsen-Freeman, 2006). During the 1970s, researchers sought to develop a developmental index for second language acquisition, aiming to "expediently and reliably gauge proficiency and development in an L2" (Larsen-Freeman, 1978, p. 440) in a quantitative, verifiable and objective way. One result of these efforts is the development of complexity, accuracy and fluency triad, generally referred to as CAF measures.

In broad terms, complexity encompasses the extent, richness, elaborateness, and diversity of L2 performance; accuracy refers to the extent to which language use is error-free and aligns with target norms; and fluency evaluates the smoothness, speed, and ease of both speech and written production. Norris and Ortega (2009) emphasize that the primary aim of measuring L2 CAF is to elucidate why and how competencies of language develop in response to specific tasks, teaching methods, and other stimuli. They argue that CAF dimensions are effective in characterizing different levels of L2 performance, thereby offering insights into developmental rates, trajectories, and ultimate outcomes. It is commonly assumed that more proficient L2 learners, or those who have undergone targeted instructional interventions, will demonstrate greater complexity in their language structures, higher accuracy, and improved fluency compared to less proficient learners or their performance prior to intervention (Pallotti, 2009).

Since their introduction into second language education, complexity, accuracy, and fluency measures have emerged as major research variables in second language acquisition (SLA) studies. These dimensions serve dual purposes: they are utilized both as performance descriptors in the spoken and written evaluation of language learners (Barrot & Gabinete, 2019; Larsen-Freeman, 2006; Şahin Kızıl, 2023) and as indicators of learners' underlying proficiency that informs their performance (Tavakoli, Kendon, Mazhurnaya, & Ziomek, 2023). From a pedagogical perspective, the CAF measures have also been found to be useful in identifying specific areas of strength and weakness in language learners, allowing educators to tailor instruction and interventions to meet individual needs effectively (Samoudi & Modirkhamene, 2020; Vercellotti, 2017; Wigglesworth & Storch, 2009).

The growing interest in employing Complexity, Accuracy, and Fluency (CAF) as research variables to evaluate learner language performance or the impact of pedagogical interventions, such as planning time or task type, has prompted a more rigorous examination of these constructs. Fundamental questions have emerged regarding the precise nature of complexity, the most effective measures of accuracy, the constituent elements of fluency, and the intricate interplay among these dimensions and their respective subcomponents (Michel, 2017).

This paper aims to provide a comprehensive overview of the CAF triad. The subsequent section presents each of these dimensions—complexity, accuracy, and fluency—along with their origins and definitions. Following this, various metrics proposed for measuring CAF are discussed in relation with findings of empirical studies that have employed CAF measures, highlighting experimental evidence relevant to second language performance. Finally, future directions in CAF research and the broader implications of CAF studies for language assessment and teaching practices are explored.

Complexity, Accuracy and Fluency: Origins and Definitions

The origins of the CAF framework can be traced back to the 70s and 80s, a period marked by a growing interest in developing quantitative and objective measures to assess L2 proficiency and development (Housen, Kuiken, & Vedder, 2012). Early attempts were inspired by the studies conducted in the field of first language acquisition for which mean length of utterance was regarded as a developmental index (Michel, 2017; Wolfe-Quintero, Inagaki, & Kim, 1998). Researchers in the field of second language were motivated by the need for reliable and valid indices that could provide a thorough picture of language learners' performance across different dimensions of language use.

One of the pioneering and most influential studies in this field is by Larsen-Freeman (1978), who emphasized the importance of developing measures that could correctly and reliably measure language proficiency and development in an L2. Concurrently, classroom-based research on L2 performance flourished during this period, with a predominant focus on characterizing language use in terms of fluency and accuracy (Brumfit, 1979). This led to the conceptualization of fluency and accuracy as key dimensions of language performance. Each dimension was identified as essential for capturing different aspects of language proficiency and for providing a holistic understanding of L2 development.

The 1980s and 1990s saw further refinement and expansion of the CAF framework, with researchers such as Skehan (1989) and Ellis (1994) contributing to the theoretical and empirical foundations of these measures. Skehan (1989), for example, proposed that complexity, accuracy and fluency are interrelated but distinct dimensions that together provide a comprehensive picture of language proficiency. Ellis (1994) emphasized the importance of balancing these dimensions in language teaching and assessment, highlighting their complementary roles in L2 acquisition.

Over the years, the CAF framework has been adopted and adapted by numerous researchers, becoming a central focus in SLA research. It has been used to explore various aspects of L2 development (e.g., Miyamoto, 2019), including the effects of instructional interventions (e.g., Samoudi & Modirkhamene, 2020), task complexity and task-specific features (e.g., Cho, 2015), and individual learner differences (e.g., Vercellotti, 2017). The CAF measures have also been applied in a wide range of contexts, from classroom-based studies to experimental research, demonstrating their versatility and relevance in the field of SLA (Barrot & Gabinete, 2019).

Despite extensive research dedicated to the CAF model, the precise definitions and operationalizations of complexity, accuracy, and fluency remain subject to ongoing debate (Phuoc & Barrot, 2022; Wolfaardt & Leung, 2023). While there is a shared understanding of these constructs within the field of second language acquisition, a closer examination reveals a diversity of interpretations and measurement approaches.

Complexity, often considered the most elusive of the CAF constructs (Pallotti, 2009), has been applied across various facets of second language acquisition research. Michelle (2017) categorizes complexity into three primary dimensions: developmental, cognitive, and linguistic. Developmental complexity pertains to the sequential emergence and mastery of linguistic structures (Pallotti, 2015). Cognitive complexity, on the other hand, reflects the subjective difficulty learners perceive in processing and acquiring specific language features. In contrast, linguistic complexity focuses on the inherent structural and semantic properties of language elements, irrespective of learner perception (Housen et al., 2012). At its core, complexity is characterised “a quality (or property) of a phenomenon or entity in terms of (1) the number and the nature of the discrete components of the entity and (2) the number and nature of the relationships between the constituent components” (Bulté & Housen, 2012, p. 22). This definition encompasses both the quantity and quality of components, as well as the nature of their connections. From a psycholinguistic perspective, complexity is conceptualized in two primary ways: absolute and relative. Absolute complexity focuses on the objective structural properties of a language system, quantifying its complexity based on the density and number of its constituent components (Bulté & Housen, 2012). In contrast, relative complexity emphasizes the cognitive demands placed on language learners, considering the mental effort required to process and produce specific linguistic features. To further refine the concept of complexity within the context of L2 performance, Bulté and Housen (2012) propose three distinct types: prepositional, discourse-interactional, and linguistic complexity. Prepositional complexity pertains to the number of ideas conveyed within a given utterance, while discourse-interactional complexity focuses on the complexity of communicative interactions. Linguistic complexity, on the other hand, encompasses both global and local dimensions. Global complexity refers to the overall sophistication of a learner's language system, whereas local complexity examines the complexity of specific linguistic features.

While complexity has presented significant challenges in terms of definitive operationalization, the constructs of accuracy, although not without their complexities, have been more readily operationalized. Accuracy is typically defined as the adherence to established grammatical and structural norms, resulting in error-free language production (Skehan & Foster, 2012). However, this seemingly straightforward concept is fraught with complexities. Defining the appropriate norm against which to measure accuracy presents a significant challenge (Michel, 2017). Should it be a prescriptive grammar, a representative corpus of native speaker language, or a dynamic standard that evolves over time? Even if a consensus were reached on the appropriate norm, determining the degree of deviation from that norm remains a significant challenge (Wolfaardt & Leung, 2023). For example, a minor grammatical error might be considered inconsequential in informal communication but deemed unacceptable in academic writing. Moreover, the concept of error severity itself is subjective, as different researchers and teachers may assign varying degrees of importance to specific errors. Kuiken and Vedder (2008) proposed a three-tiered system to categorize errors based on their impact on communicative adequacy, ranging from minor to severe. However, even this system requires judgment and interpretation. More recently, Foster and Wigglesworth (2016) introduced a weighted measure of accuracy, assigning scores to clauses based on their correctness. While offering a quantitative approach, this method also relies on subjective judgments in determining the weight assigned to each error. As Foster and Wigglesworth acknowledge, human judgment remains indispensable in the assessment of accuracy, highlighting the complexities involved in developing reliable and valid measures. This further complicates the accurate measurement of accuracy, highlighting the need for nuanced approaches to error analysis and interpretation. Moreover, the relationship between accuracy and communicative effectiveness is intricate. While grammatical correctness is essential, it does not guarantee effective communication. The concept of acceptability, encompassing both grammaticality and appropriateness, offers a more holistic perspective on language use. To further complicate matters, the measurement of accuracy is influenced by the specific linguistic features under scrutiny. For instance, assessing lexical accuracy differs markedly from evaluating grammatical accuracy. This variability necessitates context-specific approaches to accuracy assessment. Despite these complexities, accuracy remains a critical component of L2 proficiency, serving as a valuable indicator of language development. However, a nuanced understanding of its multifaceted nature is essential for effective research and pedagogy.

Finally, fluency as the third component of CAF triad is defined as the learners' ability to produce language appropriately in a given time (Wolfe-Quintero et al., 1998). While commonly associated with speech, it also extends to written language production for which it is operationalized referring to the total number of words per text (Gebril & Plakans, 2013). Early definitions emphasized speed and accuracy, equating fluency with rapid, error-free output. However, subsequent research has revealed a more nuanced understanding of the phenomenon. Distinguishing between cognitive and performance aspect of fluency, Segalowitz (2010) proposed a tripartite model of L2 fluency encompassing utterance, cognitive and perceived fluency. Cognitive fluency pertains to the underlying mental processes involved in language production, such as lexical retrieval and syntactic planning. Utterance fluency, on the other hand, refers to observable features of speech, including speech rate, pausing, and intonation. Finally, perceived fluency represents listeners' subjective judgments about a speaker's fluency based on observable characteristics. This distinction highlights the complexity of fluency, and the challenges associated with its measurement. While objective metrics, such as speech rate and pause duration, can provide valuable insights, they do not fully capture the subjective experience of fluency. Additionally, the relationship between utterance and cognitive fluency is not always straightforward. For example, a speaker may exhibit high levels of utterance fluency but experience cognitive difficulties in accessing and retrieving lexical items. Extant measures of fluency often rely on temporal aspects of speech, such as speech rate and pause duration (Chambers, 1997). While these metrics provide valuable insights, it is essential to acknowledge that fluency is a complex construct influenced by both cognitive and performance factors (de Jong et al., 2015). Alternative approaches, such as phonation time ratio (de Jong et al., 2012), offer more nuanced perspectives on fluency by accounting for the proportion of time spent actively producing speech. Additionally, the relationship between fluency and other linguistic dimensions, like complexity and accuracy, is not always straightforward (Kormos & Dénes, 2004). Measuring fluency in written language presents unique challenges due to the iterative nature of the writing process. Traditional metrics, such as word count and clause length, offer limited insights into the dynamic nature of writing fluency. However, recent advancements in the field have enabled researchers to capture more fine-grained data on writing processes, facilitating a deeper understanding of fluency in this modality.

Overall, the CAF framework offers a valuable lens for investigating the multifaceted nature of L2 performance. By examining complexity, accuracy, and fluency, researchers gain insights into learners' linguistic development. However, the operationalization of these constructs presents significant challenges due to their intricate nature. A comprehensive understanding of the CAF model is essential for effective measurement indices. The subsequent section provides an overview of the metrics commonly employed to quantify these dimensions in L2 research.

CAF Measures in SLA Research

Measuring Complexity

A closer examination of previous studies on language development and proficiency, as well as studies assessing the effect of various interventions on learners' language output, reveals a consistent focus on various metrics of Complexity, Accuracy, and Fluency (CAF). These metrics are employed in diverse forms across different studies. Wolfe-Quintero et al. (1998) categorize these measures into three main types: (a) frequency-based counts of specific linguistic units, such as the number of word tokens; (b) proportional measures, which divide one certain unit by the total number of another unit, such as the type-token ratio (TTR); and (c) indices that are calculated using more complex formulas. These varied approaches underline the complex nature of language performance and the necessity for a nuanced understanding of how different metrics can provide insights into different aspects of language proficiency.

Choosing appropriate measures of complexity, accuracy, and fluency requires careful consideration. This section does not aim to identify the 'best' measures but instead offers an overview of the commonly used metrics along with recent studies that provide reflections on these selected metrics. The intent is to guide researchers in understanding the diverse range of available measures and their application in various contexts. Table 1 presents the metrics used to measure complexity in L2 performance. It is important to note that no distinction has been made between spoken and written performance in the presentation of these metrics. However, it is acknowledged that investigating written and spoken performance may necessitate different considerations when deciding on the appropriate metric. The nuances of each modality can influence which measures are most suitable and how they should be interpreted, underscoring the importance of context-specific analysis in SLA research.

Table 1. Common measures for assessing complexity

Dimension	Type of Measure/Index	Code	Sample Study
Overall complexity	Number of Clauses Number of Verb phrases Number of T-units Number of complex T-units Clauses per sentence	C VP T CT C/S	(Xu, 2023)
Length of production unit	Mean length of sentence Mean length of clause Mean length of T-unit	MLS MLC MLT	(Lu, 2010)
Amount of subordination	Clauses per T-unit Complex clauses per T-unit Dependent clauses per clause Dependent clauses per clause	C/T CT/T DC/C DC/T	
Amount of coordination	Coordinate phrases per clause Coordinate phrases per T-unit T-units per sentence	CP/C CP/T T/S	
Degree of phrasal sophistication	Complex nominals per clause Complex nominals per T-unit Verb phrases per T-unit	CN/C CN/T VP/T	
Noun phrase density	mean number of modifiers per noun phrase	M/NP	(Tabari & Hui, 2024)
Lexical Complexity	Mean word length Lexical frequency profile Mean length of sequential word strings in a text that maintain a given TTR value Mean length of sequential lemma strings in a text that maintain a given TTR value	MTLD word MTLD lemma	(Kisselev, Soyan, Pastushenkov, & Merrill, 2022)

As illustrated in Table 1, complexity in language proficiency is assessed through various dimensions. One dimension is overall complexity, which encompasses several metrics such as the number of clauses, verb phrases (VPs), T-units (T), and complex T-units (CT). Additionally, overall complexity is evaluated by measuring clauses per sentence (C/S). Research has shown positive correlations between these overall complexity metrics and proficiency levels. For instance, Xu (2023) demonstrated that an increase in the number of clauses per sentence was linked to higher proficiency levels in learners, suggesting that more proficient writers use more complex sentence structures. However, it is important to acknowledge counterarguments in the literature. Gaillat, Lafontaine and Knefati (2023) reported that only metrics related to subordination predicted proficiency, while the number of T-units did not exhibit significant predictive power for overall proficiency. Another important dimension of complexity is the length of production which is evaluated through metrics such as the mean length of T-unit (MLT), mean length of clause (MLC) and mean length of sentence (MLS). The Mean Length of Clause (MLC) assesses the complexity of individual clauses; longer clauses generally contain more detailed information and exhibit greater structural variation, indicative of a higher level of linguistic proficiency. Similarly, the Mean Length of T-unit (MLT) correlates with proficiency levels, as more advanced learners tend to produce longer and more intricate T-units. Both MLC and MLT offer valuable insights into the sophistication of a learner's syntactic production, with longer measures often associated with increased proficiency and complexity in language use. Lu (2010) found that more proficient L2 writers tend to produce longer sentences, suggesting that MLS is a reliable indicator of syntactic complexity and developmental progress. Further support for the significance of length of production metrics is provided by Barrot and Agdeppa (2021), who confirm that indices related to the length of production units (MLS, MLT, and MLC) effectively differentiate between proficiency levels and show a linear progression across varying levels of proficiency. The amount of subordination and coordination, assessed through metrics such as DC/C, C/T, CT/T and DC/T for subordination, and CP/T, CP/C and T/S for coordination, has been the subject of various studies. Research indicates that these metrics reveal distinct patterns in syntactic complexity across different proficiency levels. Metrics related to subordination—specifically CT/T, DC/C, and DC/T—demonstrate a linear progression with increasing proficiency, reflecting a consistent advancement in the use of complex sentence structures. Conversely, the amount of coordination exhibits more variable patterns. For instance, T/S shows a steady increase in the early stages but levels off at higher proficiency levels. CP/T progresses linearly in the early stages but tends to stabilize in more advanced stages. In contrast, CP/C remains constant initially, increases at an intermediate stage, and then plateaus as proficiency advances. Overall sentence complexity (C/S) increases linearly in the early stages but declines at more advanced levels (Barrot & Agdeppa, 2021). These findings suggest that while subordination complexity develops predictably with proficiency implying a better performance, the use of coordination reflects a more varied trajectory, highlighting different developmental paths in syntactic complexity. Finally, both phrasal and lexical complexity measures are integral aspects of overall language complexity. Among the metrics commonly used in the literature, CN/C and CN/T assess the use of complex nominal structures, such as noun phrases with multiple modifiers, with higher values indicating a learner's ability to produce detailed and sophisticated descriptions. Lu (2011) found that more proficient tertiary-level language learners frequently used complex nominal phrases in their written production. Similarly, Sarte and Gnevsheva (2022) analysed 64 argumentative essays and found that learners with lower proficiency used less noun modifiers compared to their counterparts with higher-proficiency. The metric VP/T, along with M/NP, further highlights advanced descriptive skills and lexical sophistication, as more proficient learners tend to use more modifiers per noun phrase (Tabari & Hui, 2024). Research yields cautionary results regarding the use of mean word length as a metric for linguistic complexity. While studies suggest that longer words generally indicate higher linguistic complexity (Kisselev et al., 2022) and that mean word length tends to increase with instructional exposure and correlates with higher writing ratings (Barkaoui & Hadidi, 2020), it is essential to consider its sensitivity to genre and task type. Evidence shows that the average number of syllables per word can predict scores by human raters for independent tasks but not for integrated tasks (Barkaoui & Hadidi, 2020). Therefore, mean word length may reflect genre and task-specific characteristics more accurately than it does better language performance.

Measuring Accuracy

Considered as one of the most transparent constructs in the CAF triad (Michel, 2017), accuracy pertains to the extent of target-like language use, measuring the degree of deviation from established linguistic norms. The challenge in assessing accuracy lies in selecting appropriate linguistic norms and addressing the complexities associated with deviations from these norms. Recent research has introduced various indices to measure accuracy, reflecting ongoing efforts to refine the evaluation of error rates and align assessments with linguistic standards. Table 2 displays the commonly reported indices for accuracy.

Table 2. Common measures for assessing accuracy

Dimension	Type of Measure/Index	Code	Sample Study
Error free performance	Ratio of error-free T-units to all T-units	EFT/T	(Şahin Kızıl, 2023)
	Ratio of error-free clauses to all clauses	EFC/C	
	Ratio of error-free AS-units to all AS-units	EFAS	(Skehan, Bui, Wang, & Shum, 2024)
Word level Accuracy	Errors per 100 words		
Clause level accuracy	The sum of clauses of length greater than 2 multiplied by number of correct clauses at that length)/Total number of clauses	AccuracyLambda	
Subordination	Ratio of clauses to all AS units	C/AS	
	Weighted clause ratio	WCR	(Barrot & Agdeppa, 2021)

Among the most commonly used accuracy metrics is the ratio of error-free clauses (EFC/C), which stands out for its simplicity and ease of application. This measure calculates the percentage of clauses that are free from errors, offering a straightforward indicator of grammatical accuracy. A related metric, errors-per-AS-unit (EFAS), assesses the proportion of error-free Analysis of Speech (AS) -units. Polio and Shea (2014) examined the validity and reliability of these metrics, revealing that both EFC/C and the ratio of error-free T-units (EFT/T) exhibit high correlations with various error measures. This suggests that these metrics can reliably reflect language accuracy. Despite their utility, critics highlight several limitations of these measures. Notably, they do not account for the severity of errors, treating all errors uniformly regardless of their impact on meaning (Kormos, 2011; Larsen-Freeman, 2006). Additionally, EFC/C and EFAS do not consider clause length, which can lead to skewed results that favour shorter, error-free clauses while disadvantaging longer, more complex ones with fewer errors. An alternative approach proposed by Foster and Wigglesworth (2016) is what is known as the Weighted Clause Ratio (WCR), which addresses the severity of errors rather than relying on a binary classification of accuracy. Unlike traditional measures that merely categorize language as correct or incorrect, WCR offers a more nuanced evaluation by identifying various levels of error severity within written output. This method provides deeper insights into students' writing performance by distinguishing between different degrees of error impact. Evans, Hartshorn, Cox, and Martin de Jel (2014) tested the WCR and found it to be exhibiting more robustness compared to alternative accuracy metrics, demonstrating its effectiveness in capturing the complexity of language use. More recently, Barrot and Agdeppa, (2021) reported that WCR as a measure of accuracy is a robust index of proficiency with a capacity to distinguish among proficiency levels. AccuracyLambda, another metric employed in various studies to assess accuracy, is calculated by multiplying the number of correct clauses of a specific length by that length, summing these values, and then dividing by the total number of correct clauses. This approach aims to provide a nuanced measure of accuracy by considering the length of clauses in the evaluation process. However, research findings indicating high correlations between Words-per-Clause and AccuracyLambda raise concerns about the metric's distinctiveness (Skehan et al., 2024). The observed strong correlation suggests that AccuracyLambda may largely reflect differences in clause length rather than a specific aspect of accuracy. Consequently, this measure ought to be used with caution, as it may not adequately capture accuracy as a standalone construct and could be influenced by the length of the clauses being analysed (Skehan et al., 2024).

Measuring Fluency

The last constituent of the CAF triad is fluency. While there is a wide consensus on the significance of fluency as a key characteristic of successful oral communication, there is considerable divergence regarding its definition, understanding, and measurement. Additionally, there is a lack of agreement on which features of speech most accurately represent fluency at various levels of proficiency. Table 3 presents the common metrics employed to measure various aspects of fluency in second language research.

Table 3. Common measures for assessing fluency

Dimension	Type of Measure/Index	Sample Study
Breakdown fluency	Average number of silent pauses of 0.5 s or more (between and within AS-units) per AS-unit Average number of filled pauses including those of less than 0.5 s (between and within AS-units) per AS-unit Silent pause ratio Filled pause ratio: i.e. total filled-pause duration divided by total speaking-time duration.	(Jabbari & Peterson, 2023)
Repair Fluency	Number of repetitions, restarts, false starts, and repairs per minute	
Speed Fluency	Phonation time ratio Articulation rate per minute Speech rate per minute Number of pauses greater than 1 second	(Tavakoli, 2016)
Dialogue only measures	Number of turns and number of interruptions	

As shown in Table 3, fluency can be divided into several measurable dimensions, each capturing different aspects of a learner's language performance. Breakdown fluency characterized by the frequency and duration of pauses within speech can be examined in various ways that account for the location, character and amount of pausing. The amount of pause can be measured using metrics such as phonation time ratio, which evaluates the proportion of time spent while speaking versus silent instances. Breakdown fluency can also be analysed in terms of the frequency or length of pauses. For instance, Bosker et al. (2013) demonstrated that pause frequency, rather than pause length, is a more critical indicator of L2 fluency breakdown. The significance of pause location (i.e., pauses occurring in the initial, middle or final positions in a clause) was initially highlighted by Tavakoli (2011). Tavakoli (2011) argued that the differentiating factor in pausing behaviour between L1 and L2 speakers lies not in the quantity but in the placement of pauses. Subsequent research by de Jong (2016) has expanded upon this notion. The central argument is that while L1 speakers primarily employ pauses during the conceptualization phase to refine their preverbal message, L2 speakers may strategically utilize mid-clause pauses for message formulation. Speed fluency, encompassing measures such as articulation rate, speech rate, and mean length of run, has been demonstrated to effectively distinguish performance levels in second language proficiency. Articulation rate pertains to the number of syllables or words used per unit of time, excluding pauses, while speech rate includes pauses and provides a broader measure of fluency. The mean length of run measures the average number of syllables or words generated between pauses. These measures have proven successful in differentiating performance at various proficiency levels. However, they did not show significant differences at more advanced levels, suggesting a potential ceiling effect where speed increases with proficiency to a certain point but does not significantly differ at higher levels (Tavakoli, Nakatsuhara, & Hunter, 2020). This highlights the nuanced role of speed fluency in language assessment and the need for more sophisticated or varied measures to capture higher proficiency distinctions. Although fluency measures were initially associated primarily with spoken performance, recent literature on CAF has increasingly applied various metrics to assess fluency in written performance. Barrot and Agdeppa (2021), for instance, employed measures such as the number of T-units per text, the number of words per text and the number of clauses per text to evaluate writing fluency. They found that W/Tx, i.e., the number of words per text is a viable index of language proficiency. Conversely, while T/Tx, i.e., the number of T-units per text and C/Tx, i.e., the number of clauses per text also differentiate proficiency levels, they do so with a declining trend. The researchers' findings indicate that more proficient L2 writers employ fewer T-units and clauses compared to their less proficient counterparts. This outcome contradicts the assertion in the previous studies reporting that higher proficiency levels correlate with increased T-unit usage (Wolfe-Quintero et al., 1998).

Future Directions in CAF Research

Despite significant advancements in Complexity, Accuracy, and Fluency (CAF) literature, several areas still warrant further investigation and refinement. One prominent area for future research is the development of non-redundant, reliable measures of CAF. Kuiken (2023) emphasizes the need to address redundancy in existing measures, a concern initially highlighted by Norris and Ortega (2009), who encourgaed researchers to test and refine CAF measures. A potential response to this call involves exploring automated measures of CAF (Kuiken, 2023). While automated tools show promise, they currently exhibit limitations. Michel (2017) emphasizes the critical role of further developing and validating computer-based tools, leveraging advancements in machine learning and natural language processing for developing reliable and scalable CAF measurement systems. Such tools have the potential to facilitate large-scale assessments and provide more consistent and objective measurements.

Additionally, there is a need for measures that can describe language performance accurately across levels of language proficiency. Previous studies have shown that at advanced proficiency levels, some measures including morphological complexity and mean length of utterance may plateau. Instead, alternative measures, such as phraseological sophistication (Paquot, 2019), may better capture advanced L2 performance. This suggests that the development of measures tailored to various proficiency stages is essential.

Longitudinal studies are also necessary to complement cross-sectional designs, which often fail to capture developmental patterns of complexity. Relevant literature highlights the variability in CAF development at the individual level, which often diverges from mean group trends. Therefore, integrating longitudinal case studies with group studies could help identify generalizable developmental patterns. Also, a cross-linguistic perspective is valuable for understanding how language complexity varies across different languages. Extant literature underlines the significance of examining the impact of L1 configurations and differences between native speakers and non-native speakers. Expanding research to include a broader range of target languages, beyond the predominant focus on English and other languages, is also necessary.

Furthermore, investigating how teachers address CAF aspects in instructional practice is crucial. Although teachers may not prioritize CAF when assessing L2 learners' performance, understanding CAF development is important for pedagogical goals (Norris & Ortega, 2009) Teaching these aspects effectively requires acknowledging that syntactic and lexical errors are part of the learning process.

Finally, exploring the potential of artificial intelligence in measuring CAF could further enhance the accuracy and efficiency of assessments. Advances in AI and machine learning could lead to the development of sophisticated models capable of assessing not only surface-level features but also deeper aspects of language proficiency, such as nuanced aspects of fluency and complexity. Moreover, AI tools can be designed to adapt and improve over time through continuous learning, potentially leading to more refined and personalized assessments. However, it is important to address ethical considerations and ensure that AI systems are transparent and free from biases. Continued research into the integration of AI with traditional assessment methods could pave the way for more robust and comprehensive evaluations of language performance and proficiency through CAF.

Conclusion

In conclusion, the CAF triad offers a comprehensive framework for assessing L2 performance and proficiency. While significant progress has been made in understanding and measuring these dimensions, challenges remain in refining metrics and exploring technological advancements. This paper has provided an overview of the CAF triad, highlighting definitions, measurement referring to empirical findings, and future research directions.

In terms of complexity, various measures such as overall complexity, length of production, and subordination versus coordination were presented. Research indicates that increased complexity in language use—evidenced by metrics like clauses per sentence and mean length of T-unit—correlates with higher proficiency levels. However, the development of complexity measures shows varied patterns across proficiency stages, with subordination exhibiting a more predictable progression compared to coordination.

For accuracy, this paper focused on metrics such as error-free clauses and error rates, noting their utility and limitations. While traditional measures like the proportion of error-free clauses (EFC/C) provide a straightforward assessment of grammatical accuracy, they may not fully account for error severity or clause length. Alternative metrics, such as the Weighted Clause Ratio (WCR) and AccuracyLambda, offer more nuanced insights but require careful interpretation due to their potential overlap with other constructs.

In examining fluency, this paper presented metrics related to breakdown, repair, and speed fluency. Measures like articulation rate and mean length of run effectively distinguish proficiency levels, though some metrics show limited differentiation at advanced stages. The application of fluency measures to written performance further extends our understanding, though results indicate that the use of T-units and clauses in writing may not align with previously established patterns for spoken performance.

Advancements in CAF research hold significant implications for both theoretical and practical aspects of second language acquisition (SLA). Firstly, refining CAF measures can lead to more precise assessments of language proficiency, which is crucial for educational and professional contexts. Enhanced CAF metrics enable educators to tailor instruction more effectively, addressing specific developmental needs and thereby improving learning outcomes and proficiency levels.

Secondly, robust CAF measures have the potential to inform language testing and certification processes. High-stakes language assessments can integrate nuanced CAF metrics to provide a more comprehensive evaluation of test-takers' abilities, ensuring that assessments are fair and reflective of actual language use. This approach benefits learners by offering clearer benchmarks and more targeted feedback on their language skills.

Finally, the development of automated tools for CAF measurement has implications for large-scale language assessment and research. Such tools can facilitate the analysis of extensive language corpora, enabling broader and more diverse studies on language proficiency and acquisition. This, in turn, can lead to more generalizable findings and contribute to evidence-based language teaching methodologies.

Ultimately, advancements in CAF research and measurement will enhance language teaching and assessment practices, leading to improved outcomes for language learners worldwide. The ongoing exploration and refinement of CAF metrics are essential for deepening our understanding of second language acquisition and improving the ways we evaluate and support language learners.

References

Barrot, J. S., Agdeppa, J. Y. (2021). Complexity, accuracy, and fluency as indices of college-level L2 writers' proficiency. *Assessing Writing*, *47* (December 2020), 100510. https://doi.org/10.1016/j.asw.2020.100510

Barkaoui, K., Hadidi, A. (2020). *Assessing changes in second language writing performance.* London: Routledge.

Barrot, J. S., Gabinete, M. K. (2019). Complexity, accuracy, and fluency in the argumentative writing of ESL and EFL learners. *IRAL - International Review of Applied Linguistics in Language Teaching*, *Online First*, 1-24. https://doi.org/10.1515/iral-2017-0012

Bosker, H. R., Pinget, A. F., Quené, H., Sanders, T., de Jong, N. H. (2013). What makes speech sound fluent? The contributions of pauses, speed and re-pairs. *Language Testing*, 30, 159-175.

Brumfit, C. (1979). Communicative language teaching: An educational perspective. In C. Brumfit & Johnson K (Eds.), *The communicative approach to language teaching* (pp. 183-191). London: Oxford University Press.

Bulté, B., Housen, A. (2012). *Defining and operationalising L2 complexity*. https://doi.org/10.1075/lllt.32.02bul

Cho, H. (2015). Effects of Task Complexity on English Argumentative Writing. *English Teaching*, *70*(2), 107-131. https://doi.org/10.15858/engtea.70.2.201506.107

de Jong, N. H. (2016a). Predicting pauses in L1 and L2speech: The effects of utterance boundaries and word frequency. *International Review of Applied Lin-guistics in Language Teaching,* 54, 113-132.

Ellis, R. (1994). *The study of second language acquisition*. Oxford: Oxford University Press.

Evans, N. W., Hartshorn, K. J., Cox, T. L., Martin de Jel, T. (2014). Measuring written linguistic accuracy with weighted clause ratios: A question of validity. *Journal of Second Language Writing*, *24*(1), 33-50. https://doi.org/10.1016/j.jslw.2014.02.005

Foster, P., Wigglesworth, G. (2016). Capturing accuracy in second language performance: The case for a weighted clause ratio. *Annual Review of Applied Linguistics*, Vol. 36, pp. 98-116. Cambridge University Press. https://doi.org/10.1017/S0267190515000082

Gaillat, T., Lafontaine, A., Knefati, A. (2023). Visualizing Linguistic Complexity and Proficiency in Learner English Writings Visualising linguistic complexity and proficiency in learner English writings. *A Knefati CALICO Journal*, *40*(2), 178-197. https://doi.org/10.1558/cj.19487ï

Gebril, A., Plakans, L. (2013). Toward a transparent construct of reading-to-write tasks: The interface between discourse features and proficiency. *Language Assessment Quarterly*, *10*(1), 9-27. https://doi.org/10.1080/15434303.2011.642040

Housen, A., Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics*, *30*(4), 461-473. https://doi.org/10.1093/applin/amp048

Housen, A., Kuiken, F., Vedder, I. (2012). Complexity, accuracy and fluency: Definitions, measurement and research. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 Performance and Proficiency* (pp. 1–21). Amsterdam / Philadelphia: John Benjamins Publishing Company. Retrieved from http://benjamins.com/catalog/lllt

Jabbari, N., Peterson, M. (2023). Complexity, accuracy, and fluency improvements through massively multiplayer online gaming: a longitudinal mixed-methods case study. *Language Learning Journal*, *51*(4), 416-450. https://doi.org/10.1080/09571736.2023.2219713

Kisselev, O., Soyan, R., Pastushenkov, D., Merrill, J. (2022). Measuring writing development and proficiency gains using indices of lexical and syntactic complexity: Evidence from longitudinal Russian learner corpus data. *Modern Language Journal*. https://doi.org/10.1111/modl.12808

Kuiken, F. (2023). Linguistic complexity in second language acquisition. *Linguistics Vanguard*, *9*(s1), 83-93. https://doi.org/10.1515/lingvan-2021-0112

Larsen-Freeman, D. (1978). An ESL Index of Development. *TESOL Quarterly*, *12*(4), 439-448. Retrieved from https://about.jstor.org/terms

Larsen-Freeman, D. (2006). The emergence of complexity, fluency, and accuracy in the oral and written production of five Chinese learners of English. *Applied Linguistics*, *27*(4), 590-619. https://doi.org/10.1093/applin/aml029

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

Michel, M. (2017). Complexity, Accuracy and Fluency (CAF). In S. Laowen & M. Sato (Eds.), *The Routledge Handbook of Instructed Second Language Acquisition* (pp. 1-38). London: Routledge.

Miyamoto, M. (2019). *Capturing L2 oral proficiency with CAF measures as predictors of the ACTEFL OPI rating* (Doctoral dissertation). Purdue University, West Lafayette.

Norris, J. M., Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics*, *30*(4), 555-578. https://doi.org/10.1093/applin/amp044

Pallotti, G. (2009). CAF: defining, refining and differentiating constructs. *Applied Linguistics*, *30*(4), 590-601.

Pallotti, G. (2015). A simple view of linguistic complexity. *Second Language Research*, 31(1), 117-134. doi:10.1177/0267658314536435

Phuoc, V. D., Barrot, J. S. (2022). Complexity, accuracy, and fluency in L2 writing across proficiency levels: A matter of L1 background? *Assessing Writing*, *54*. https://doi.org/10.1016/j.asw.2022.100673

Şahin Kızıl, A. (2023). Data-driven learning: English as a foreign language writing and complexity, accuracy and fluency measures. *Journal of Computer Assisted Learning*, *39*(4), 1382-1395. https://doi.org/10.1111/jcal.12807

Samoudi, N., Modirkhamene, S. (2020). Concordancing in writing pedagogy and CAF measures of writing. *IRAL - International Review of Applied Linguistics in Language Teaching*. https://doi.org/10.1515/iral-2020-2014

Sarte, K. M., Gnevsheva, K. (2022). Noun phrasal complexity in ESL written essays under a constructed-response task: Examining proficiency and topic effects. *Assessing Writing*, 51. https://doi.org/10.1016/j.asw.2021.100595

Segalowitz, N. (2010). *Cognitive bases of second language fluency.* London: Routledge.

Skehan, P. (1989). *Individual differences in second language learning*. London: Edward Arnold.

Skehan, P., Bui, G., Wang, Z., Shum, S. (2024). Re-examining accuracy measures in second language task-based spoken performance. *Research Methods in Applied Linguistics*, *3*(1). https://doi.org/10.1016/j.rmal.2024.100098

Skehan, P., Foster, P. (2012). Complexity, accuracy, fluency and lexis in task-based performance: a synthesis of the Ealing research. *Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA.*, (1987), 199-220. https://doi.org/10.1075/lllt.32.09fos

Tabari, M. A., Hui, B. (2024). Exploring the associations among task complexity, task motivation, task engagement, and linguistic complexity in L2 writing. *Modern Language Journal*, *108*(2), 446-468. https://doi.org/10.1111/modl.12921

Tavakoli, P. (2011). Pausing patterns: Differences be-tween L2 learners and native speakers. *ELT Journal*, 65, 71-79.

Tavakoli, P. (2016). Fluency in monologic and dialogic task performance: challenges in defining and measuring L2 fluency. *International Review of Applied Linguistics in Language Teaching*, *54*(2), 133-150. https://doi.org/https://doi.org/10.1515/iral-2016-9994

Tavakoli, P., Kendon, G., Mazhurnaya, S., Ziomek, A. (2023). Assessment of fluency in the Test of English for Educational Purposes. *Language Testing*, *40*(3), 607-629. https://doi.org/10.1177/02655322231151384

Tavakoli, P., Nakatsuhara, F., Hunter, A. –M. (2020). Aspects of Fluency Across Assessed Levels of Speaking Proficiency. *Modern Language Journal*, *104*(1), 169-191. https://doi.org/10.1111/modl.12620

Vercellotti, M. Lou. (2017). The development of complexity, accuracy, and fluency in second language performance: A longitudinal study. *Applied Linguistics*, *38*(1), 90-111. https://doi.org/10.1093/applin/amv002

Wigglesworth, G., Storch, N. (2009). Pair versus individual writing: Effects on fluency, complexity and accuracy. *Language Testing*, *26*(3), 445-466. https://doi.org/10.1177/0265532209104670

Wolfaardt, J. F., Leung, A. H.-C. (2023). The Impact of Filipina Domestic Workers on Hong Kong Primary School Children's L2 English Spoken CAF and Reading Accuracy and Fluency. *Applied Linguistics*. https://doi.org/10.1093/applin/amad072

Wolfe-Quintero, K., Inagaki, S., Kim, H.-Y. (1998). *Second language development in writing: measures of fluency, accuracy, and complexity*. Honolulu: University of Hawai'i Press. https://doi.org/10.1017/S0272263101263050

Xu, P. (2023). Reconsidering the syntactic complexity measures on L2 spoken English: A multi-dimensional perspective. *Heliyon*, *9*(6). https://doi.org/10.1016/j.heliyon.2023.e16856