



Castledown

 OPEN ACCESS

Language Education & Assessment

ISSN 2209-3591

<https://www.castledown.com/journals/lea/>

Language Education & Assessment, 9, 105239 (2026)

<https://doi.org/10.29140/lea.2026.105239>

Validating Automated Complexity, Accuracy, and Fluency Metrics for L2 Writing Analysis



THOMAS DILLON 

Daegu Catholic University, South Korea

lucalicious@gmail.com

Abstract

This study evaluates the construct validity of automated complexity, accuracy, and fluency (CAF) metrics for AI-assisted analysis of L2 writing. As automated feedback systems and learner analytics increasingly rely on computational indicators to evaluate learner texts at scale, it is necessary to establish whether these metrics support defensible interpretations before pedagogical or longitudinal use. Using a cross-sectional dataset of 1,000 Korean EFL learner texts and minimally corrected counterparts generated by a fine-tuned grammatical error correction model, we examined candidate CAF indicators within a structural equation modeling framework. Accuracy was operationalized using GLEU and edit/error count, fluency using raw perplexity (PPL) and delta perplexity (Δ PPL), and complexity using selected L2SCA indices. GLEU and edit/error count converged but were redundant, supporting GLEU as a single observed accuracy indicator. Raw PPL and Δ PPL also converged in isolation, but their latent fluency specification became inadmissible in the full model. For complexity, a three-indicator factor comprising MLT, MLC, and CN/C produced a stable measurement structure. In the final admissible SEM, Δ PPL showed stronger covariance with GLEU, whereas raw PPL showed stronger covariance with Complexity, indicating a trade-off between correction-normalized fluency and construct separation. The findings support a reduced framework while highlighting the need to compare PPL and Δ PPL further in longitudinal modeling.

Keywords: Contemporary validity theory, CAF measures, assessment integrity

Copyright: © 2026 Thomas Dillon. This is an open access article distributed under the terms of the Creative Commons Attribution Non-Commercial 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within this paper.

Introduction

The Complexity–Accuracy–Fluency (CAF) framework models second language (L2) performance as a multidimensional construct and is widely used to compare proficiency, evaluate instruction, and track development in L2 writing (Housen, Kuiken, & Vedder, 2012; Norris & Ortega, 2009; Ortega, 2003). As learner corpora have expanded, CAF research has increasingly relied on automated pipelines, including the L2 Syntactic Complexity Analyzer (L2SCA) for syntactic indices and grammatical error correction (GEC) metrics such as GLEU for correction-based accuracy estimates (Lu, 2011; Napoles, Sakaguchi, Post, & Tetreault, 2016). This shift improves scalability but raises a central measurement problem: whether computational proxies represent the theoretical constructs they are intended to measure (Biber, Gray, Staples, & Egbert, 2020).

This problem is especially acute because CAF indicators can be heterogeneous, redundant, or dependent on pipeline choices. Omnibus measures such as Mean Length of T-unit (MLT) collapse clausal subordination and phrasal elaboration into a single value, making them difficult to interpret linguistically; at higher proficiency levels, clausal subordination plateaus while phrasal compression and noun-phrase density better capture advanced academic writing development (Norris & Ortega, 2009; Biber, Gray, & Poonpon, 2011; Biber & Gray, 2016; Biber et al., 2020). Accuracy measures derived from GEC systems introduce a parallel problem: reference-based metrics often penalize valid corrections absent from gold-standard references, while automated correction outputs embed a correction policy, such as minimal repair or stylistic rewriting, that can shift the construct being measured (Bryant, Felice, & Briscoe, 2017; Dahlmeier & Ng, 2012; Napoles, Sakaguchi, & Tetreault, 2016, 2017; Sumner & Dillon, 2026). Recent work has questioned the reliability of traditional n-gram metrics for evaluating modern neural corrections and has proposed LLM-based or multi-reference alternatives to better account for legitimate linguistic variation (Kobayashi, Mita, & Komachi, 2024; Goto, Sakai, & Watanabe, 2025; Klinger et al., 2025).

Under contemporary validity theory, validity is not an intrinsic property of a metric but a function of score interpretation and use (Messick, 1994; Kane, 2013; Chapelle & Voss, 2013). For CAF analysis, this requires evidence that indicators relate in ways consistent with their proposed constructs and exhibit a stable internal structure (Kline, 2011; Vafae & Kachinske, 2019). This study evaluates automated indicators of accuracy, fluency, and complexity: GLEU and edit-based measures for accuracy, language-model perplexity (PPL) and Δ PPL for fluency, and L2SCA indices of length and phrasal sophistication for complexity (Lu, 2011; Bryant et al., 2017; Barrot & Agdeppa, 2021; Sánchez, Dobnik, & Volodina, 2024; Sumner & Dillon, 2026). Using Confirmatory Factor Analysis (CFA), tests whether the indicators provide internal-structure evidence consistent with separable CAF dimensions, as predicted by CAF theory, rather than collapsing into a single proficiency factor (Koizumi & In'nami, 2014; Kim et al., 2018; Vafae & Kachinske, 2019). Establishing this structure supports subsequent longitudinal modeling of L2 writing development and broader applications in computer-assisted feedback, learner analytics, and progress monitoring across instructional contexts (Barkaoui, 2014; Murakami, 2016; Vercellotti, 2017).

Literature Review

The CAF Framework: Theoretical Foundations and Tensions

The Complexity–Accuracy–Fluency (CAF) framework in SLA models L2 performance as multidimensional, not unitary, based on the assumption that development occurs through partially independent, potentially asynchronous dimensions (Housen et al., 2012; Pallotti, 2015). In L2 writing, complexity denotes the elaboration and sophistication of grammatical (and often lexical) resources, accuracy denotes conformity to target-language norms, and fluency denotes ease, naturalness, and processing efficiency in language use (Ortega, 2003; Thewissen, 2013; Segalowitz, 2010; Vercellotti, 2017).

A central claim in CAF theory is that these dimensions compete under limited attentional resources. Skehan's Limited Capacity Hypothesis predicts trade-offs, whereby increased attention to structural elaboration can reduce resources available for accuracy monitoring or fluent production (Skehan, 1998). By contrast, Robinson's Cognition Hypothesis predicts that increased task complexity can promote simultaneous gains in complexity and accuracy, though often with fluency costs under greater processing demands (Robinson, 2011; Plakans et al., 2019). Recent evidence from computer-mediated environments supports these cognitive tensions, demonstrating that while automated written corrective feedback can significantly improve accuracy, it often fails to promote growth in syntactic complexity or fluency, as learners selectively prioritize grammatical repair over other dimensions of performance (Cheng & Zhang, 2025).

Empirical findings do not support a single stable relationship among CAF dimensions. Accuracy–complexity relations vary by proficiency and task, and can be competitive, facilitative, or sequential rather than uniformly trade-off based. Covariation among indicators is expected, but does not remove the need to test whether they form distinguishable dimensions in a given dataset (Thewissen & Anishchanka, 2022).

This issue is partly methodological. Norris and Ortega (2009) argue that metric choice can determine whether indicators behave as construct-specific measures or as proxies for general proficiency. Consistent with this, some studies report separable latent constructs in writing data, whereas others report correlations large enough to suggest functional redundancy (Koizumi & In'nami, 2014). The issue is not the validity of CAF theory itself, but whether measurement choices preserve multidimensionality in a given dataset.

Automation intensifies this problem by expanding index availability while weakening interpretability. Omnibus or length-based measures often collapse distinct forms of syntactic elaboration into a single score, obscuring whether higher values reflect clausal subordination, phrasal elaboration, or verbosity (Biber et al., 2020). Recent AI-mediated writing research reinforces this concern from both assessment and feedback perspectives: fine-tuned generative models often outperform traditional linguistic complexity measures in predicting human ratings, yet still vary across L1 groups and lower-proficiency essays, while AI- and teacher-mediated corrective feedback can improve accuracy and writing quality without producing parallel gains in syntactic complexity (Liu, Qi, & Lu, 2025; Abdi Tabari, Kushki, & Wang, 2025). In longitudinal research, this is consequential because apparent fluctuation can reflect metric instability, learner-group bias, feedback effects, or construct contamination rather than development (Larsen-Freeman, 2006; Boulton & Vyatkina, 2021; Murakami, 2016). Automated CAF analysis requires evidence that selected indicators preserve the theoretical distinction among complexity, accuracy, and fluency. Establishing this structure is a prerequisite for later growth modeling and for interpreting trade-offs or co-development as developmental effects (Koizumi & In'nami, 2014; Barkaoui, 2014; Murakami, 2016).

Operationalizing Complexity

A persistent problem in CAF research is that complexity is not unitary, but a set of partially overlapping constructs whose behavior varies across tasks, genres, and proficiency levels. Consequently, relations between syntactic complexity and proficiency depend on both how proficiency is defined and which metrics are used (Ortega, 2003). Omnibus indices such as mean unit length or global subordination ratios are convenient but linguistically coarse, often obscuring distinct developmental patterns and ignoring the phraseological dimension of complexity, including sophisticated word combinations and “complex linguistic gestalts” (Norris & Ortega, 2009; Bulté & Housen, 2014; Paquot, 2019).

A major shift in the literature is the claim that advanced academic writing develops increasingly through phrasal compression, especially dense noun-phrase elaboration, rather than continued growth

in clausal subordination. On this account, clausal measures alone underrepresent advanced writing because syntactic sophistication is realized through informational density within clauses rather than additional subordination. This helps explain why subordination indices plateau or decline at higher proficiency levels, while phraseological and phrasal sophistication continue to distinguish advanced development (Biber et al., 2011; Biber & Gray, 2016; Paquot, 2019).

This issue is especially important in computational CAF research, where automated pipelines often default to length and subordination metrics that are easy to compute but theoretically underspecified. Lu's work established the feasibility of automated syntactic complexity measurement and showed that indices differ in developmental sensitivity (Lu, 2010; Lu, 2011). Later studies suggest that fine-grained clausal and phrasal measures can outperform traditional length or subordination indices in predicting writing quality, although features associated with holistic quality do not necessarily track instructional development, creating a construct–criterion mismatch for validation work (Crossley & McNamara, 2014; Kyle & Crossley, 2018; Crossley, 2020).

The present study operationalizes syntactic complexity using a parsimonious subset of L2SCA indices. Although L2SCA provides 14 measures, not all are equally stable or discriminative across proficiency levels (Lu, 2010; Lu, 2011; Lei et al., 2023). We prioritize theoretically complementary and empirically supported indicators: length-based measures such as MLT and MLC as broad markers of elaboration, and phrasal measures such as CN/C and, where appropriate, CN/T as indicators of noun-phrase density and informational packaging (Lu, 2011; Biber et al., 2011; Biber & Gray, 2016; Barrot & Agdeppa, 2021). This treats complexity not as a single index, but as a multidimensional construct whose indicators should show interpretable shared variance consistent with theory and recent factor-analytic work (Biber et al., 2020; Lan, Li, & Zhang, 2022).

Operationalizing Accuracy

In L2 writing research, accuracy has traditionally been operationalized through manual error analysis, including error-free clause or T-unit ratios and coded error categories representing conformity to target-language norms (Wolfe-Quintero et al., 1998; Housen et al., 2012). This foundational work established accuracy as a relatively stable and internally coherent construct, typically defined as the degree of conformity to target-language norms (Polio, 1997; Pallotti, 2015). As learner corpora have expanded, automated accuracy measurement has increasingly relied on NLP and grammatical error correction (GEC) metrics, especially similarity- and edit-based measures comparing learner texts with corrected versions (Napoles et al., 2016; Bryant et al., 2017). This improves scalability but changes the construct: accuracy is no longer measured directly as a property of the learner text, but as distance from a reference correction.

This reference dependence creates a validity issue because similarity-based scores measure distance to a correction policy rather than correctness in the abstract. Different correction systems prioritize minimal repair, stylistic rewriting, simplification, or fluency-oriented paraphrase, yielding different accuracy scores for the same learner text (Napoles et al., 2016; Bryant et al., 2023). This study uses a profile of complementary indicators rather than assuming that any single metric is universally superior (Polio & Shea, 2014). GLEU measures n-gram overlap between the learner text and corrected reference while penalizing uncorrected error-related n-grams, providing a bounded similarity measure suitable for comparison across texts (Napoles et al., 2016). ERRANT-derived edit counts provide a more transparent estimate of the amount of modification required to reach the reference norm (Bryant et al., 2017). Together, these indicators capture both similarity to the minimally corrected benchmark and the volume of required correction.

A central risk is that GEC systems are not neutral measurement instruments. Contemporary neural and LLM-based correctors often produce paraphrases, substitutions, or broader fluency-oriented rewrites beyond a minimal-edit objective, causing overlap- and edit-based scores to reflect stylistic divergence rather than grammatical error density alone (Katinskaia & Yangarber, 2024; Bryant et al., 2023; Sumner & Dillon, 2026). Accuracy operationalization requires explicit validity justification: the key question is whether GLEU and edit burden behave as coherent accuracy indicators and remain empirically separable from fluency within the measurement model. Testing their clustering and discriminant behavior in CFA/SEM serves as a safeguard against construct conflation introduced by the reference problem (Koizumi & In'nami, 2014; Kane, 2013).

Operationalizing Fluency

Writing fluency refers to the efficiency, automaticity, and naturalness of producing connected text. Because direct process measures such as pausing behavior or keystroke logging are impractical in longitudinal online research, written fluency is often inferred from product-based indicators that approximate well-formedness rather than real-time production speed (Segalowitz, 2010; Kim et al., 2018; Latif, 2013; Plakans et al., 2019; Vercellotti, 2017).

Language-model perplexity (PPL) is one such proxy and is best interpreted as an index of distributional naturalness rather than cognitive fluency itself. PPL quantifies how “surprising” a text is to a probabilistic language model; lower values indicate greater predictability and stronger conformity to target-language regularities. However, raw PPL is noisy in learner corpora because it is sensitive to factors beyond proficiency, including lexical rarity, topic, proper nouns, specialized vocabulary, punctuation, and task-specific phrasing (Islam & Magnani, 2021; Yang et al., 2012; Kharkwal & Muresan, 2013; Sánchez et al., 2024; Serrano et al., 2023; Wense, Rothe, & Filippova, 2018; Liu, 2020).

To address this limitation, the present study also evaluates Δ Perplexity: $\Delta\text{PPL} = \text{PPL}_{\text{source}} - \text{PPL}_{\text{corrected}}$

Because the source and minimally corrected texts share topic, lexical content, and intended meaning, ΔPPL functions as a within-text normalization measure. It may reduce topic- and vocabulary-related variance while preserving sensitivity to correction-related improvements in naturalness and grammatical well-formedness. It also avoids dependence on external references, which are vulnerable to low-coverage bias because the space of valid corrections cannot be fully represented by limited “gold standard” references (Islam & Magnani, 2021; Sumner & Dillon, 2026; Wense et al., 2018; Serrano et al., 2023; Choshen & Abend, 2018).

However, ΔPPL is not an absolute fluency measure. Because it is derived from the source–correction relationship, it reflects correction-normalized change and may share variance with correction-referenced accuracy indicators such as GLEU or edit burden. Raw PPL and ΔPPL are treated as competing fluency-related operationalizations: raw PPL captures absolute distributional well-formedness, whereas ΔPPL captures the reduction in surprisal after minimal correction. The CFA/SEM framework tests which indicator provides the more defensible fluency measure and whether fluency remains empirically distinguishable from accuracy and complexity (Sumner & Dillon, 2026; Choshen & Abend, 2018; Kim et al., 2018; Koizumi & In'nami, 2014; Kane, 2013).

The Argument-Based Validation Frame

This study adopts an argument-based view of validity, in which validity depends not on the metric itself but on the interpretation and use of scores within a specified evidentiary argument (Messick, 1994; Kane, 2013; AERA, APA, & NCME, 2014). For CAF measurement, the core claim is that

NLP-derived indicators can be interpreted as evidence of complexity, accuracy, and fluency rather than as proxies for general proficiency, topic variation, or pipeline artifacts.

Internal structure provides the key evidence for this claim. Confirmatory Factor Analysis (CFA) tests whether the observed covariance among indicators is consistent with the proposed CAF structure and allows comparison with rival models, such as a single proficiency factor or models in which accuracy and fluency collapse because both depend on correction outputs (Koizumi & In'nami, 2014; Kim et al., 2018; Kline, 2023). This comparison is necessary because acceptable fit for one model does not, by itself, establish that the proposed construct interpretation is uniquely warranted.

Such evidence is essential before longitudinal modeling. If retained indicators reflect measurement noise, correction-policy artifacts, or undifferentiated proficiency rather than distinct CAF dimensions, later claims about trade-offs, co-development, or nonlinear growth would be weakly supported. Establishing separable CAF dimensions in the present dataset provides the measurement basis for interpreting subsequent developmental trajectories (Barkaoui, 2014; Murakami, 2016).

Synthesis, Gap, and Research Questions

The literature reviewed above shows that, although automated analysis is now feasible at scale, the interpretability of automated scores depends on whether selected indicators form admissible, stable, and empirically differentiated measurement structures within a given learner corpus. At the same time, GLEU and perplexity-based metrics have been used mainly in GEC and writing-evaluation research rather than validated as components of a unified CAF measurement system, leaving unclear whether they can function as defensible indicators of Accuracy and Fluency alongside syntactic complexity indices. The present study uses a two-stage SEM/CFA design to evaluate candidate indicators for admissibility and redundancy and then to test whether retained Accuracy, Fluency, and Complexity measures remain empirically differentiated, including comparison of raw PPL and Δ PPL as alternative fluency operationalizations.

This yields two research questions:

- RQ1.** Which candidate accuracy, fluency, and complexity metrics form admissible measurement structures for the framework?
- RQ2.** To what extent are the retained Accuracy, Fluency, and Complexity measures empirically differentiated in the full SEM, and how do raw PPL and Δ PPL differ as observed fluency indicators?

Methodology Overview: SEM-Based Validation Strategy

The aim of this study is methodological: to determine whether a set of automated indices can support a defensible CAF interpretation in this dataset and, if so, to identify a measurement structure suitable for subsequent longitudinal modeling. We used SEM as a pragmatic validation framework in two stages: first, we screened candidate indicators for Accuracy, Fluency, and Complexity to determine which could be retained as admissible components of the measurement system; second, we evaluated the retained structure in a full SEM and compared alternative observed fluency indicators.

Data and Candidate Variables

Analyses were conducted on a cross-sectional sample of 1,000 texts with complete data for the candidate CAF indicators. Accuracy was represented by corrected_GLEU and an edit/error proxy (Total_errors) for screening purposes. Fluency was represented by perplexity of the original text (text_ppl) and a derived fluency index, Δ PPL ($\text{ppl_delta} = \text{text_ppl} - \text{corrected_ppl}$). Complexity was initially examined

using four L2SCA indices representing complementary aspects of syntactic complexity: MLT, MLC, CN_C, and CN_T. All variables were z-standardized prior to model estimation. An anonymized extract illustrating the structure of the analysis dataset is provided in Appendix D.

Descriptive Statistics

Descriptive screening showed that multiple indicators were non-normal, with expected strong right skew for the perplexity measures. This distributional pattern is not unexpected for perplexity, which is lower-bounded and tends to cluster at relatively low values while a smaller number of texts receive disproportionately high scores. The strongest distributional irregularity was observed for text_MLT, which showed extreme positive skew and kurtosis. Based on follow-up inspection of the most extreme cases, this pattern likely reflects instability in automated T-unit segmentation for weakly punctuated learner texts rather than greater syntactic complexity.

Table 1 summarizes the descriptive statistics for the candidate CAF indicators across 1,000 learner texts. The reported means, standard deviations, skewness, and kurtosis provide an overview of their distributions before measurement-model screening.

Table 1 *Descriptive statistics for candidate CAF indicators*

Variable	Mean	SD	Skew	Kurtosis
GLEU	75.42	12.07	−0.64	0.49
Total_errors	15.18	8.73	0.99	1.27
text_ppl	46.76	25.21	1.99	6.04
corrected_ppl	24.67	11.09	3.04	17.63
ppl_delta	22.10	19.09	2.21	7.91
text_MLT	12.82	6.28	9.03	129.16
text_MLC	8.05	2.02	2.34	10.87
text_CN_C	0.71	0.36	2.13	9.28
text_CN_T	1.16	0.83	5.69	65.99

Note: n = 1000 for all variables. Full table available in Appendix C.

Screening Candidate Measurement Structures

Because some automated indices were expected to overlap substantially, we first evaluated whether candidate indicators could function as stable latent structures. For Accuracy and Fluency, this involved two-indicator latent “pair tests” estimated in isolation:

$$\text{Accuracy} = \sim \text{corrected_GLEU} + \text{Total_errors}$$

$$\text{Fluency} = \sim \text{text_ppl} + \text{ppl_delta}$$

For each pair test, we examined (a) the correlation between indicators, (b) standardized loadings, and (c) indicator residual variances. These checks were used to identify either redundancy (very small residual variance indicating near-duplication) or inadmissibility (negative residual variance or non-convergence).

Complexity was examined through confirmatory models comparing a retained 3-indicator structure (MLT, MLC, CN_C) with an expanded 4-indicator structure that additionally included CN_T. This comparison was used to determine whether the broader candidate set improved representation of syntactic complexity or instead introduced instability.

System-level Admissibility Screening

Two-indicator factors can appear acceptable in isolation but become unstable when embedded in a broader SEM. We conducted an additional system-level screening model in which latent Fluency (PPL + Δ PPL) was entered alongside latent Complexity and covariances with observed Accuracy. This served as a stress test of admissibility under the intended SEM context. If this model produced an inadmissible solution (e.g., a Heywood case for Δ PPL), Fluency was treated as observed in the retained SEM structure.

Because raw PPL and Δ PPL capture partly different aspects of fluency-related variance—absolute well-formedness versus correction-normalized change—both were retained for comparison at the observed-variable level if the latent Fluency specification proved inadmissible.

Retained SEM Structure and Parallel Fluency Comparison

Following the screening stage, we specified a retained admissible SEM intended to balance statistical and theoretical adequacy. This structure treated:

Complexity as a latent factor indicated by MLT, MLC, and CN_C;

Accuracy as an observed indicator represented by corrected_GLEU; and

Fluency as an observed indicator in the baseline model, represented by Δ PPL.

Rather than imposing directional developmental hypotheses, the SEM estimated covariances among the retained Accuracy, Fluency, and Complexity measures, consistent with the study's cross-sectional validation aim. Model fit was evaluated using standard global fit indices (χ^2 , CFI, TLI, RMSEA, SRMR) and parameter estimates (standardized loadings, covariances, and residual variances).

To evaluate whether the fluency results were specific to the use of Δ PPL, we also estimated a parallel admissible SEM replacing Δ PPL with raw PPL, as well as a diagnostic joint model including both PPL and Δ PPL. These supplementary models were used to compare how the two fluency indicators covaried with Accuracy and Complexity. The R code used for indicator screening, Heywood-case diagnostics, and estimation of the final and parallel SEMs is provided in Appendix E.

Decision Rule for Retaining Latent Versus Observed Indicators

Where a candidate latent factor produced inadmissible or unstable solutions in the broader SEM context (e.g., negative indicator residual variance or nonconvergence), the construct was retained using an observed indicator or reduced indicator set that remained both interpretable and statistically admissible. This rule was intended to ensure that the final framework was suitable for the CAF pipeline.

When multiple admissible observed indicators were available, retention was not determined solely by global model fit. We also considered discriminant behavior with the other CAF dimensions. In particular, raw PPL and Δ PPL were compared with respect to their covariance with observed Accuracy (GLEU) and the latent Complexity factor in the parallel SEMs.

Results

This section reports the results in CAF order: Complexity, Accuracy, and Fluency. The text beneath each table explains what the reported values show in descriptive terms only, focusing on convergence, loadings, residual variances, model fit, and covariances.

RQ1. Which Candidate Complexity, Accuracy, and Fluency Metrics Formed Admissible Measurement Structures for the Framework?

Table 2 summarizes the screening decisions for the candidate CAF measurement structures. Detailed evidence for each retained or excluded structure is reported in the following subsections.

Table 2 *Screening summary of candidate CAF measurement structures*

Model	Indicators	Conv.	Inadm.	Key evidence	Retained
Complexity 3-ind.	MLT, MLC, CN_C	Yes	No	Stable loadings; positive residuals	Yes
Complexity 4-ind.	MLT, MLC, CN_C, CN_T	No	Yes	CN_T destabilized model	No
Accuracy pair	GLEU, errors	Yes	No	Near-redundant inverse pair	No
Fluency pair	PPL, Δ PPL	Yes	No	Strong convergence in isolation	No
Fluency in SEM	PPL, Δ PPL	Yes	Yes	Joint latent specification inadmissible	No

Complexity

Complexity was screened first. The initial 3-indicator solution and the expanded 4-indicator solution are reported separately in Tables 2 and 3.

Table 3 *Complexity screening: 3-indicator model*

Indicator	Std. loading	Residual variance (Est.)	Residual variance (Std.)
text_MLT	0.391	0.847	0.847
text_MLC	0.820	0.327	0.327
text_CN_C	0.900	0.190	0.190

Note: Converged = Yes. Heywood / inadmissible = No.

Table 3 shows that all three retained indicators loaded positively on the latent Complexity factor and all residual variances remained positive. The standardized loadings indicate that MLC and CN_C contributed the largest share of common variance, while MLT contributed additional but smaller shared variance.

Table 4 *Complexity screening: expanded 4-indicator model including CN/T*

Indicator	Std. loading	Residual variance (Est.)	Residual variance (Std.)
text_MLT	0.028	0.998	0.999
text_MLC	−0.014	0.999	1.000
text_CN_C	0.025	0.998	0.999
text_CN_T	24.860	−746.290	−617.028

Note: Converged = No. Heywood / inadmissible = Yes.

Table 4 reports the attempted expansion of the Complexity factor by adding CN_T. This model did not converge and produced inadmissible estimates, including an extreme standardized loading for CN_T and a negative residual variance. These values indicate that the 4-indicator specification could not be retained.

Accuracy

Accuracy was screened as a two-indicator pair consisting of corrected_GLEU and Total_errors. The purpose of this step was to determine whether the two indicators behaved as a coherent accuracy-related set and whether they contributed distinct information.

Table 5 *Accuracy screening: two-indicator latent pair test*

Indicator	Std. loading	Residual variance (Est.)	Residual variance (Std.)
corrected_GLEU	0.975	0.049	0.049
Total_errors	-0.972	0.056	0.056

Note: Pair correlation $r = -0.948$. Converged = Yes. Heywood / inadmissible = No.

Table 5 shows that the two Accuracy indicators were strongly related, as reflected in the large inverse correlation and the large standardized loadings of opposite sign. The small positive residual variances indicate that little variance remained outside the shared accuracy-related signal, so the pair functioned as a near-redundant measurement set at the screening stage.

Fluency

Fluency was screened in two steps. First, raw PPL and Δ PPL were tested together as a two-indicator pair in isolation. Second, that joint latent specification was checked again inside the broader SEM context.

Table 6 *Fluency screening: two-indicator latent pair test*

Indicator	Std. loading	Residual variance (Est.)	Residual variance (Std.)
text_ppl	0.948	0.101	0.101
ppl_delta	0.961	0.076	0.076

Note: Pair correlation $r = 0.911$. Converged = Yes. Heywood / inadmissible = No.

Table 6 shows that raw PPL and Δ PPL converged strongly in the pair test. Both indicators had high standardized loadings and both residual variances remained positive, indicating that the pair behaved acceptably when estimated in isolation.

Table 7 *System-level fluency screening: residual variances in latent-Fluency SEM*

Indicator	Residual variance (Est.)
text_ppl	0.313
ppl_delta	-0.208

Note: Converged = Yes. Heywood / inadmissible = Yes. A negative residual variance indicates a Heywood case.

Table 7 reports the same Fluency indicators after they were placed inside the broader SEM. In this setting, the joint latent Fluency specification became inadmissible because one residual variance became negative. This result indicates the two indicators could not be retained together as a stable latent Fluency factor in the full model, so they were carried forward as alternative observed specifications under RQ2.

RQ2. To What Extent were the Retained Complexity, Accuracy, and Fluency Measures Empirically Differentiated in the Full SEM, and How did Raw PPL and Δ PPL Differ as Observed Fluency Indicators?

After the screening stage, the retained SEM structure kept Complexity as a latent factor, used corrected_GLEU as the observed Accuracy indicator, and compared two observed Fluency specifications: a baseline model using Δ PPL and a parallel model using raw PPL. The tables below report model fit and the covariances among the retained measures.

Table 8 *Comparison of admissible full-SEM models using Δ PPL and PPL as observed Fluency indicators*

Model	CFI	TLI	RMSEA	SRMR	Complexity ~~ Fluency (Std.)	GLEU ~~ Fluency (Std.)
Δ PPL observed	0.999	0.997	0.022	0.016	0.140	-0.699
PPL observed	0.997	0.992	0.034	0.019	0.339	-0.531

Table 8 compares the two admissible SEMs. Both models showed good global fit. The table also gives the standardized covariances most relevant to the fluency comparison: the association between Complexity and the Fluency indicator, and the association between GLEU and the Fluency indicator.

Table 9 *Baseline final SEM covariances (Accuracy = GLEU, Fluency = Δ PPL)*

Covariance	Estimate	SE	z	p	Std.
Complexity ~~ ppl_delta	0.140	0.034	4.141	< .001	0.140
Complexity ~~ corrected_GLEU	0.052	0.034	1.545	.122	0.052
ppl_delta ~~ corrected_GLEU	-0.698	0.039	-18.111	< .001	-0.699

Table 9 reports the covariances for the baseline final SEM using Δ PPL. The standardized covariance column shows that Complexity had only a small positive association with Δ PPL and a very small, non-significant association with GLEU, whereas Δ PPL and GLEU were strongly negatively related in this specification.

Table 10 *Parallel SEM covariances (Accuracy = GLEU, Fluency = PPL)*

Covariance	Estimate	SE	z	p	Std.
Complexity ~~ text_ppl	0.339	0.033	10.279	< .001	0.339
Complexity ~~ corrected_GLEU	0.054	0.034	1.585	.113	0.054
text_ppl ~~ corrected_GLEU	-0.530	0.036	-14.823	< .001	-0.531

Table 10 reports the same covariance structure after Δ PPL was replaced with raw PPL. In this parallel model, the standardized covariance between Fluency and GLEU remained negative but was smaller in magnitude than in the Δ PPL model, while the standardized covariance between Fluency and Complexity was larger.

Taken together, the RQ2 tables show that the retained CAF measures did not collapse into a single undifferentiated pattern. They also show that changing the observed Fluency specification changed the size of the covariances involving Fluency, while the Complexity–Accuracy relation remained small in both admissible models.

Discussion

From a CAF perspective, interrelationships among dimensions are expected rather than problematic. Limited-capacity and trade-off accounts propose that learners allocate finite attentional resources across competing performance demands, producing negative or complex associations among complexity, accuracy, and fluency depending on task conditions and proficiency (e.g., Skehan, 1998). By contrast, cognition-oriented accounts predict that under some conditions dimensions co-develop rather than trade off, particularly when tasks or instructional support promote simultaneous engagement with form and meaning (e.g., Robinson, 2011). The present results are best interpreted within that tradition: the aim is not to force orthogonality among CAF dimensions, but to identify an admissible automated measurement structure that can support later longitudinal modeling.

This issue is especially relevant for CALL because automated feedback systems and learner analytics increasingly transform learner writing into computational indicators that may inform instruction, progress monitoring, and research decisions. Without validation, such indicators risk becoming technically convenient but pedagogically under-specified measures.

The first task of the study was measurement selection: to determine which indicators could be retained in the framework without redundancy, over-parameterization, or inadmissible solutions. The second task was structural interpretation: to determine whether the retained measures behaved as empirically differentiated CAF dimensions in the full SEM, and what the comparison between raw PPL and Δ PPL implies for the next longitudinal phase.

RQ1: Which Candidate Complexity, Accuracy, and Fluency Metrics Formed Admissible Measurement Structures for the Framework?

For Accuracy, corrected_GLEU and Total_errors showed strong convergence, indicating that both metrics are capturing the same underlying correction-distance signal. Texts that are closer to their minimally corrected versions receive higher GLEU scores and, correspondingly, require fewer edits. This pattern provides useful convergent evidence, but it also indicates substantial redundancy. This does not mean that either metric is invalid; modeling both together added little unique information.

Within that pair, GLEU is the stronger candidate for retention in the final model for methodological reasons. Unlike Total_errors, which is a raw integer count and partly dependent on essay length and the granularity of individual edit operations, GLEU is easier to compare across texts and to interpret directionally: higher values reflect the extent of overlap between the original and corrected text. Total_errors remains useful as corroborating screening evidence, but GLEU is better suited as the accuracy index retained in the SEM.

This clarifies claims about accuracy. The present study does not argue that GLEU is a gold-standard measure of absolute correctness in the abstract. Rather, it shows that within this automated pipeline

GLEU behaves coherently with an error-burden proxy and can function as a defensible observed indicator of correction-referenced accuracy. Because the corrected benchmark is generated by a previously validated minimal-edit model, the resulting score is interpretable as distance to a pedagogically aligned correction policy rather than as an unqualified statement of linguistic correctness. For the purposes of large-scale automation, that is a sufficient measurement claim.

At the screening level for Fluency, raw PPL and Δ PPL clearly converged: they shared a strong association and behaved as related fluency-oriented indicators. When both were brought together as a latent Fluency factor inside the broader SEM, the joint specification was inadmissible. Similarly to Accuracy metrics, there is no gain in forcing both indicators into the same latent fluency measurement model within the full CAF system.

This distinction matters because the parallel SEM comparison suggests that both PPL and Δ PPL remain analytically informative. Δ PPL has a conceptual appeal in this project because it normalizes within text, reducing topic and content noise by comparing the original essay against its own corrected counterpart. At the same time, that normalization also makes it more tightly bound to the correction pipeline and more liable to overlap with the correction-referenced Accuracy measure. Raw PPL behaves differently: it is less directly anchored to the source-correction comparison, but also absorbs broader variance associated with distributional well-formedness, lexical profile, and general language development. The implication is not that one metric should be discarded at the screening stage. Rather, the choice between them affects how Fluency relates to the rest of the CAF system, so both are examined in the parallel SEMs.

For Complexity, the retained three-indicator solution proved stable and theoretically coherent. The selected measures captured complementary aspects of syntactic elaboration without producing the identification problems that emerged in the expanded solution. This supports a parsimonious approach to automated syntactic complexity: the aim is not to include every available L2SCA measure, but to retain a subset that preserves interpretability while avoiding multicollinearity and construct dilution. In this dataset, the retained indicators supported a latent Complexity factor and justified carrying it forward as the only latent component in the final CAF specification. This factor is a restricted operationalization of syntactic complexity: its cross-sectional coherence does not establish sensitivity to developmental change, and it may obscure change in individual indicators or other syntactic dimensions.

RQ2: To What Extent were the Retained Complexity, Accuracy, and Fluency Measures Empirically Differentiated in the Full SEM, and How did Raw PPL and Δ PPL Differ as Observed Fluency Indicators?

The full SEM results indicate that the retained CAF dimensions were neither fully independent nor collapsed into a single undifferentiated proficiency construct. Complexity remained only weakly related to Accuracy, which supports its interpretation as a distinct dimension of learner production rather than a by-product of correction-based well-formedness. This is important because it shows that the selected L2SCA indicators are contributing information that is not reducible to the same signal captured by the correction-derived measures.

The more difficult interpretive issue concerns Fluency. In the baseline admissible model using Δ PPL, the covariance between Fluency and Accuracy was very strong, which is theoretically understandable because both measures are anchored to the same source-to-correction relationship. Essays that are already close to their minimally corrected versions tend to receive higher GLEU scores and also show smaller perplexity reductions after correction. Essays that require more repair show the opposite pattern. This strong coupling should not be treated as evidence that CAF has collapsed. Instead,

it suggests that Δ PPL captures a specific correction-referenced aspect of fluency: the degree to which model-based well-formedness improves under minimal repair.

The parallel model with raw PPL suggests a different trade-off. Relative to Δ PPL, raw PPL appears less tightly bound to the correction pipeline and less dominated by the same reparability signal that drives GLEU. At the same time, it shows a stronger positive relationship with Complexity, implying that it captures a broader distributional profile of learner writing that overlaps more with structural elaboration or overall language development. This does not make raw PPL preferable, any more than Δ PPL's tighter normalization makes it superior. The choice of fluency indicator changes the pattern of construct separation, and that change is meaningful.

The framework is viable, but the fluency dimension remains sensitive to operationalization. Δ PPL is a reasonable baseline for the validation study because it fits the original design logic and performed well in the admissible final model, yet the parallel PPL model shows that alternative operationalizations produce meaningfully different structural relations. The longitudinal study should compare raw PPL and Δ PPL rather than assume in advance that within-text normalization is universally the better solution.

Taken together, the findings support a practical CAF framework: Complexity can be modeled as a stable latent construct; Accuracy can be represented by GLEU; and Fluency can be modeled with either PPL or Δ PPL as an observed index. This cross-sectional validation step does not settle every theoretical question, but it does establish an admissible measurement foundation for the forthcoming longitudinal analysis of learner development at scale.

Limitations

Several limitations define the scope of the present validity argument. First, GLEU and Δ PPL depend on a single correction system and its embedded minimal-edit policy. Although the correction model and generation procedure are documented and were evaluated in our previous GEC study, these measures remain conditional on that policy and may behave differently under alternative correction systems. Because both measures use the corrected text, their relationship cannot be interpreted as wholly independent evidence of distinct constructs.

Second, the validation is cross-sectional. The results provide evidence concerning the internal structure of the selected measures, but do not establish whether those measures capture theoretically interpretable within-learner development or remain stable across stages of learning.

Third, the sample was drawn from a single L1 population of Korean EFL learners in one online writing environment. Under an argument-based view of validity, this defines the domain for which the present interpretation is supported rather than establishing a universal measurement model. Although the corpus includes heterogeneous self-selected topics and writing purposes, applying the measurement specification to other genres, tasks, proficiency distributions, or L1 populations would require further validation.

Finally, the retained Complexity factor represents shared variance among selected L2SCA syntactic indices, not the full range of writing complexity. It does not encompass lexical, discourse, or rhetorical complexity, and a coherent factor does not imply that its constituent indices will follow identical developmental trajectories. Likewise, PPL and Δ PPL are product-based distributional measures. They do not capture process dimensions of written fluency such as production speed, pausing, or revision behaviour. These boundaries cannot be removed by the present statistical analysis and should be considered when interpreting and applying the retained measures.

Future Work

The retained measurement specification will be applied to the full longitudinal corpus to examine whether the cross-sectional findings support interpretable within-learner trajectories. The analysis will compare source PPL and Δ PPL as alternative product-based fluency indicators, assessing whether they yield similar developmental patterns and whether Δ PPL's reliance on a corrected counterpart affects its relationship with Accuracy.

The longitudinal analysis will also examine the latent Complexity factor alongside its constituent indices. If the factor provides a limited summary of change, the analysis will extend to the broader set of L2SCA indices to identify which dimensions of syntactic complexity develop and whether their trajectories diverge. This comparison will test whether the latent factor retained from the cross-sectional validation is sufficiently sensitive to within-learner change to serve as a measure of syntactic complexity in longitudinal analysis.

Further research should test the robustness of these interpretations across learner populations, corpora, correction systems, genres, and tasks. Combining process-tracing methods with completed-text measures could also clarify how product-based indicators relate to writing speed, pausing, and revision behaviour.

Conclusion

This study validated an automated CAF measurement framework designed for scalable analysis of L2 writing. Using argument-based validation logic and CFA/SEM internal-structure evidence, we showed that (i) GLEU and edit/error burden provide strongly convergent—and largely redundant—accuracy information, supporting the use of a single observed accuracy index; (ii) raw PPL and Δ PPL converge at the indicator level but do not support a stable two-indicator latent Fluency factor in the full SEM, indicating that fluency remains sensitive to operationalization; and (iii) selected L2SCA indices form a stable latent Complexity construct. The resulting admissible specification provides the foundation for longitudinal modeling of CAF development at scale and for broader applications in computer-assisted feedback, learner analytics, and progress monitoring across instructional contexts, while making explicit the methodological constraints introduced by reference-based and model-based computational proxies.

Use of Generative AI

This manuscript made use of Google Gemini to assist in paraphrasing and improving the clarity of text. No AI tools were used for the generation of research hypotheses or the writing of conclusions. The final manuscript was authored and reviewed by the listed authors without additional AI input.

Funding

This research was supported by the Daegu Catholic University Research Fund.

Author Biography

Thomas Dillon is an assistant professor in the Department of English Language and Literature at Daegu Catholic University, Republic of Korea. His research focuses on computer-assisted language learning, automated writing evaluation, grammatical error correction, large language models in language education, and the longitudinal development of second language writing.

References

- Abdel Latif, M. M. M. (2013). What do we mean by writing fluency and how can it be validly measured? *Applied Linguistics*, 34(1), 99–105. <https://doi.org/10.1093/applin/ams073>
- Abdi Tabari, M., Kushki, A., & Wang, Y. (2025). Comparing the effects of teacher-and AI-mediated corrective feedback on accuracy, complexity, and quality in L2 written narratives. *Computer Assisted Language Learning*, 1–24.
- AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*. American Educational Research Association.
- Barkaoui, K. (2014). Quantitative approaches for analyzing longitudinal data in second language research. *Annual Review of Applied Linguistics*, 34, 65–101.
- Barrot, J. S., & Agdeppa, A. G. (2021). Complexity, accuracy, and fluency as indices of college-level L2 writers' proficiency. *Assessing Writing*, 47.
- Biber, D., & Gray, B. (2016). *Grammatical complexity in academic English: Linguistic change in writing*. Cambridge University Press.
- Biber, D., Gray, B., & Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly*, 45(1), 5–35. <https://doi.org/10.5054/tq.2011.244483>
- Biber, D., Gray, B., Staples, S., & Egbert, J. (2020). Investigating grammatical complexity in L2 English writing research: Linguistic description versus predictive measurement. *Journal of English for Academic Purposes*, 46, 100869. <https://doi.org/10.1016/j.jeap.2020.100869>
- Boulton, A., & Vyatkina, N. (2021). Thirty years of data-driven learning: Taking stock and charting new directions over time. *Language Learning & Technology*, 25(3), 66–89. <https://hdl.handle.net/10125/73450>
- Bryant, C., Felice, M., & Briscoe, T. (2017). Automatic annotation and evaluation of error types for grammatical error correction. In Proceedings of ACL 2017 (pp. 793–805). *Association for Computational Linguistics*. <https://doi.org/10.18653/v1/P17-1074>
- Bryant, C., Yuan, Z., Qorib, M. R., Cao, H., Ng, H. T., & Briscoe, T. (2023). Grammatical error correction: A survey of the state of the art. *Computational Linguistics*, 49(3), 643–701. https://doi.org/10.1162/coli_a_00478
- Bulté, B., & Housen, A. (2014). Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of Second Language Writing*, 26, 42–65. <https://doi.org/10.1016/j.jslw.2014.09.005>
- Chapelle, C. A., & Voss, E. (2013). Evaluation of language tests through validation research. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 1079–1097). Wiley. <https://doi.org/10.1002/9781118411360.wbcla110>
- Cheng, X., & Zhang, L. J. (2025). Investigating synchronous and asynchronous written corrective feedback in a computer-assisted environment: EFL learners' linguistic performance and perspectives. *Computer Assisted Language Learning*, 38(7), 1687–1716.
- Choshen, L., & Abend, O. (2018). Inherent biases in reference-based evaluation for grammatical error correction. In Proceedings of ACL 2018 (pp. 632–642). Association for Computational Linguistics. <https://doi.org/10.18653/v1/P18-1059>
- Crossley, S. A. (2020). Linguistic features in writing quality and development: An overview. *Journal of Writing Research*, 11(3), 415–443. <https://doi.org/10.17239/jowr-2020.11.03.01>
- Crossley, S. A., & McNamara, D. S. (2014). Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners. *Journal of Second Language Writing*, 26, 66–79. <https://doi.org/10.1016/j.jslw.2014.09.006>
- Dahlmeier, D., & Ng, H. T. (2012). Better evaluation for grammatical error correction. In Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: human language technologies (pp. 568–572).

- Goto, T., Sakai, Y., & Watanabe, T. (2025, September). Reliability crisis of reference-free metrics for grammatical error correction. In *Findings of the Association for Computational Linguistics: EMNLP 2025* (pp. 24913–24926).
- Housen, A., Kuiken, F., & Vedder, I. (2012). *Dimensions of L2 performance and proficiency*. John Benjamins.
- Islam, M. A., & Magnani, E. (2021). Is this the end of the gold standard? A straightforward reference-less grammatical error correction metric. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 3009–3015).
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73. <https://doi.org/10.1111/jedm.12000>
- Katinskaia, A., & Yangarber, R. (2024). GPT-3.5 for Grammatical Error Correction. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. (pp. 7831–7843). ELRA and ICCL. <https://aclanthology.org/2024.lrec-main.692>
- Kharkwal, G., & Muresan, S. (2014). *Surprisal as a predictor of essay quality*. In *Proceedings of BEA 2014* (pp. 54–60). Association for Computational Linguistics.
- Kim, Y. S. G., Gatlin, B., Al Otaiba, S., & Wanzek, J. (2018). Theorization and an empirical investigation of the component-based and developmental text writing fluency construct. *Journal of learning disabilities*, 51(4), 320–335.
- Klinger, E., Huang, Z., Nguyen, T. M., Park, E. J., Chen, Y., Gu, Y., ... & Park, J. (2025). A formal framework for fluency-based multi-reference evaluation in grammatical error correction. arXiv preprint arXiv:2510.06749.
- Kline, R. B. (2023). *Principles and practice of structural equation modeling*. Guilford publications.
- Kobayashi, M., Mita, M., & Komachi, M. (2024, June). Large language models are state-of-the-art evaluator for grammatical error correction. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)* (pp. 68–77).
- Koizumi, R., & In'nami, Y. (2014). Modeling complexity, accuracy, and fluency of Japanese learners of English: A structural equation modeling approach. *JALT Journal*, 36(1), 25–46. <https://doi.org/10.37546/JALTJJ36.1-2>
- Kyle, K., & Crossley, S. A. (2018). Measuring syntactic complexity in L2 writing using fine-grained clausal and phrasal indices. *The Modern Language Journal*, 102, 333–349. <https://doi.org/10.1111/modl.12468>
- Lan, G., Li, Y., & Zhang, J. (2022). Factor structure of grammatical complexity in L2 writing: A multidimensional perspective. In *Proceedings of the Workshop on Second Language Acquisition and Natural Language Processing. Association for Computational Linguistics*.
- 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.
- Lei, L., Wen, J., & Yang, X. (2023). A large-scale longitudinal study of syntactic complexity development in EFL writing: A mixed-effects model approach. *Journal of Second Language Writing*, 59, Article 100962.
- Liu, Y. (2020). *Assessing text readability and quality with language models* [Master's thesis, University of Helsinki]. HELDA. <https://hdl.handle.net/10138/313475>
- Liu, Y., Qi, H., & Lu, X. (2025). Enhancing GPT-based automated essay scoring: the impact of fine-tuning and linguistic complexity measures. *Computer Assisted Language Learning*, 1–20.
- Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15(4), 474–496.
- Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly*, 45(1), 36–62. <https://doi.org/10.5054/tq.2011.240859>

- Messick, S. J. (1994). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *ETS Research Report Series*. <https://doi.org/10.1002/j.2333-8504.1994.tb01618.x>
- Murakami, A. (2016). Modeling systematicity and individuality in nonlinear second language development: The case of English grammatical morphemes. *Language Learning*, 66(4), 834–871.
- Napoles, C., Sakaguchi, K., Post, M., & Tetreault, J. (2016). *GLEU without tuning*. arXiv. <https://arxiv.org/abs/1605.02592>
- Napoles, C., Sakaguchi, K., & Tetreault, J. (2017). JFLEG: A fluency corpus and benchmark for grammatical error correction. In *Proceedings of EACL 2017* (pp. 229–234). Association for Computational Linguistics.
- 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>
- Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics*, 24(4), 492–518. <https://doi.org/10.1093/applin/24.4.492>
- Pallotti, G. (2015). A simple view of linguistic complexity. *Second Language Research*, 31(1), 117–134.
- Paquot, M. (2019). The phraseological dimension of complexity. The phraseological dimension in interlanguage complexity research. *Second Language Research*, 35(1), 121–145.
- Plakans, L., Gebril, A., & Bilki, Z. (2019). Shaping a score: Complexity, accuracy, and fluency in integrated writing performances. *Language Testing*, 36(2), 161–179.
- Polio, C. G. (1997). Measures of linguistic accuracy in second language writing research. *Language Learning*, 47(1), 101–143.
- Polio, C., & Shea, M. C. (2014). An investigation into current measures of linguistic accuracy in second language writing research. *Journal of Second Language Writing*, 26, 10–27.
- Robinson, P. (2011). Second language task complexity, the Cognition Hypothesis, language learning, and performance. In P. Robinson (Ed.), *Second language task complexity: Researching the Cognition Hypothesis of language learning and performance* (pp. 3–38). John Benjamins. <https://doi.org/10.1075/tblt.2.05ch1>
- Sánchez, R. M., Dobnik, S., & Volodina, E. (2024). Harnessing GPT to study second language learner essays: Can we use perplexity to determine linguistic competence? In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)* (pp. 414–427).
- Segalowitz, N. (2010). *Cognitive bases of second language fluency*. Routledge.
- Serrano, S., Brumbaugh, Z., & Smith, N. A. (2023). *Language models: A guide for the perplexed*. arXiv. <https://arxiv.org/pdf/2311.17301>
- Skehan, P. (1998). *A cognitive approach to language learning*. Oxford University Press.
- Sumner, J. P., & Dillon, T. (2026). Assessing fine-tuning and instruction-prompting methods in GPT-driven grammatical error correction. *Language Education & Assessment*, 9, 103750.
- Thewissen, J. (2013). Capturing L2 accuracy developmental patterns: Insights from an error-tagged EFL learner corpus. *The Modern Language Journal*, 97(S1), 77–101.
- Thewissen, J., Anishchanka, A., Leńko-Szymańska, A., & Götz, S. (2022). Interaction between grammatical accuracy and syntactic complexity at different proficiency levels. *Complexity, Accuracy, and Fluency in Learner Corpus Research*, 209–240.
- Vafaei, P., & Kachinske, I. (2019). The inadequate use of confirmatory factor analysis in second language acquisition validation studies. *Studies in Applied Linguistics and TESOL*, 19(2).
- Vercellotti, M. L. (2017). The development of complexity, accuracy, and fluency in second language performance: A longitudinal study. *Applied linguistics*, 38(1), 90–111.

- von der Wense, K., Rothe, S., & Filippova, K. (2018). Sentence-level fluency evaluation: References help, but can be spared! In *Proceedings of the 22nd Conference on Computational Natural Language Learning* (pp. 313–323).
- Wolfe-Quintero, K., Inagaki, S., & Kim, H.-Y. (1998). *Second language development in writing: Measures of fluency, accuracy, and complexity*. University of Hawai'i Press.
- Yang, M. C., Kim, M. J., Lee, H. G., & Rim, H. C. (2012). Assessing writing fluency of non-English-speaking students for automated essay scoring: How to automatically evaluate the fluency in English essay. In *International Conference on Computer Supported Education* (Vol. 2, pp. 83–87). SCITEPRESS.

Appendix

Please see the appendices at: <https://github.com/Professor77/CAF-Validity/blob/main/Appendix>