Measuring complexity, accuracy, fluency (CAF)

Background

The acquisition and use of a second language have always been seen as multidimensional processes in SLA research. However, it was only in the mid-1990s that the Complexity, Accuracy and Fluency (CAF) triad became established as a standard way of describing the multidimensionality of language performance, in particular after the works of Skehan (1998) and Wolfe-Quintero, Inagaki and Kim (1998).

In the following years, some contributions described how these dimensions vary over time in L2 development. The bulk of research, however, focused on how the three constructs vary according to task's features, testing in particular the predictions of two theoretical models, the Cognition Hypothesis (Robinson, 2015) and the Limited Attentional Capacity Model (Skehan, 1998; 2015). Results, however, are far from being clear and consistent. The main problem lies in the fact that all these studies have used a large variety of measures and operationalizations of both independent variables (task conditions) and dependent variables (CAF measures), often paying little attention to their construct validity (Ellis, 2018; Plonsky & Kim, 2016). Only more recently some critical reviews have aimed to clarify the constructs' theoretical and operational definitions (Bulté & Housen, 2012; Lambert & Kormos, 2014; Norris & Ortega, 2009; Pallotti, 2009, 2015; Skehan & Foster, 2012).

In language testing and assessment, too, the three dimensions of Complexity, Accuracy and Fluency play an important role. Most language tests consider accuracy as one of the dimensions to be assessed, in terms of the presence of errors, while "fluency has been a component of speaking assessment since the Second World War" (Fulcher, 2015, p. 77). Complexity is also often included in performance descriptors and several studies have found a positive correlation between test scores and more complex language use (e.g. Thirakunkovit et al., 2019 and references therein).

Before turning to a presentation of the three dimensions, some general clarifications are in order. The first concerns an asymmetry between the three constructs. Complexity and Fluency have to do with the interlanguage system as such, of which they describe several characteristics in positive terms, such as the presence of some structures, or its delivery rate. Accuracy, on the other hand, implies the comparison with another system, namely the target language, which is conducted in negative terms by counting errors (cases in which certain structures are *not* produced).

A second point concerns modality. Complexity and accuracy may be evaluated in basically the same way in oral and written texts: the most important difference lies in the choice of the reference units, which may be identified in writing by means of punctuation (sentences or T-Units; Hunt, 1965), while in oral production this identification often relies on prosody and intonation (utterances or AS-Units; Foster, Tonkyn & Wigglesworth, 2000). If written data are collected with keylogging programs, even the fluency of written production can be evaluated in much the same way as in oral speech. In the absence of these direct measures, writing fluency becomes a rather different construct.

In the following pages, I will present the three dimensions, discussing, for each of them, the construct's theoretical definition and possible operationalizations. Since it is not possible to review all the measures, I will list only the most used, which have best withstood criticisms and are, in my opinion, more valid. Validity will be assessed mainly in logical-conceptual terms, as the correspondence of the measure to the construct it intends to measure, and in terms of reliability (constancy of the measure across evaluators). To my knowledge, no studies have explicitly addressed the issue of validating CAF measures, although some factor analyses have shown that the three dimensions are in fact relatively independent (e.g., Koizumi & In'nami, 2014; Norris & Ortega, 2009). Several authors (e.g., Polio & Shea, 2014; Wolfe-Quintero, Inagaki & Kim, 1998) have proposed to evaluate the validity of CAF measures based on their ability to vary over time or in relation to changes in task conditions. However, this approach is questionable, because it assumes what needs to be demonstrated: if a measure does not vary over time, or across tasks, this can simply be a fact, not a problem with the measure's validity. In other words, considering good measures only those varying where we expect variation means assuming that our theories, or expectations, are valid in themselves, which seems to be too strong an assumption in the current state of knowledge. One can speak at

most about the *practical usefulness* of a measure in discriminating learners at different levels, or for assessing the effects of different communicative activities (see e.g. Hsieh & Wang, 2019).

Key Concepts

Complexity: the number of elements and their interrelationships in a text or linguistic system. It can be grammatical (i.e. required by the linguistic system) or stylistic (i.e. the result of free choices by language users).

Accuracy: the conformity of linguistic performance to target-language norms.

Fluency: the extent to which linguistic production is (and/or perceived as) fast and smooth. Three dimensions are usually identified: speed (articulation rate), breakdown (number and length of pauses), repair (number of reformulations, false-starts etc).

Complexity and accuracy may be applied to specific levels of linguistic description: phonology, lexicon, morphosyntax, pragmatics. Fluency is more holistic and refers to language performance as a whole.

Key issues

Complexity

Learning a new language consists first of all in building a system of patterns and rules, which becomes increasingly vast and thus more complex. The term *complexity* is commonly used to refer both to a structural dimension (number of elements of a system, intricacy of their relationships), and to a cognitive dimension, as the effort needed to process the whole system or some of its structures. In the interest of terminological transparency, I would recommend reserving the term *complexity* only to the first, structural, sense and to call *difficulty* the second aspect, which is also in line with language testing parlance (Pallotti, 2009, 2015). Therefore, a text containing many "sophisticated" linguistic structures (because they are acquired late, or because they are often produced with errors) should be said to be more "difficult" or "advanced", but not more "complex". Of course, it may be empirically demonstrated that more structurally complex linguistic units tend to be more difficult, and produced at more advanced levels, but this is not a good reason to call these different constructs by the same name.

Lexical complexity

Assessing lexical complexity from a structural point of view implies calculating the variety of lexemes within a text. Since the type/token ratio (the number of times a lexical item is repeated in a text) is influenced by text length, several alternative measures have been proposed, such as the Guiraud index (types/sq root tokens; which only offers a partial correction), the Standardized Type/Token Ratio (STTR; the calculation of the average lexical diversity within samples of constant length), or MTLD, vocd-D and HD-D (for a review, McCarthy & Jarvis, 2013).

Other measures that have been proposed for assessing complexity, but actually concern difficulty, calculate the proportion of more or less rare words in a text, i.e. those belonging to the first, second, third, etc. band in a frequency ranking. The underlying idea is that if a learner produces rare words, her lexical competence may be deemed to be more advanced than that of one who is limited to a basic lexicon. All this makes sense, but I think it is more appropriate to say that a rare word like *tar* is more difficult or advanced than *car*, rather than saying it is more complex.

Morphological complexity

The few existing measures of morphological complexity gauge the average diversity of exponents, or morphological processes, in a text. Pallotti (2015; Brezina & Pallotti, 2019) proposed the Morphological

Complexity Index (MCI), which calculates the average diversity of inflectional exponents within and between samples of 10 forms randomly extracted from a text. De Clercq and Housen (2019) additionally used the Types/Family (Horst & Collins, 2006) and the Inflectional Diversity (Malvern et al., 2004) measures, that are based on the calculation of the ratio of different inflected word forms per lemma.

Syntactic complexity

In many studies, "linguistic complexity" has been represented by syntactic complexity alone and operationalized by just one or two measures, such as the average length of units or the subordination ratio - a clear example of construct under-representation (Bulté & Housen, 2012). On the other hand, many studies employed redundant measures, i.e. tapping into the same sub-dimensions of the construct, e.g. subordination ratio together with number of clauses per T-Unit (Norris & Ortega, 2009).

Unlike lexical and morphological complexity measures, which look at the variety of forms, syntactic complexity measures mainly concern the average "depth" of syntactic structures. With a view to simplification, Pallotti (2015) proposes to select only three of the many measures in this area, namely the average number of words per phrase, number of phrases per clause and number of clauses per higher-order syntactic unit (sentence, T-Unit, AS-unit). De Clercq and Housen (2017) have proposed a measure accounting for the variety of syntactic structures, too.

Fluency

Fluency can be seen as an observable characteristic of texts ("utterance fluency"), as a reflection of cognitive processes ("cognitive fluency") and as the perception of an external observer ("perceived fluency") (Segalowitz, 2010). From a cognitive perspective, fluency is connected to the dimension of "control" of language production processes – that is, with their proceduralization and automatization – and with functions such as planning and monitoring. On a descriptive level, three main dimensions are usually identified: speed, breakdown and repair (Tavakoli & Skehan, 2005).

Speed fluency. This dimension represents how fast language is produced. It may be measured by counting the number of syllables per unit of time, for example 1 minute (speech rate). However, this measure is spurious, as it confounds articulation speed and the amount of pausing. A better operationalization is articulation rate, that is, the number of syllables per speaking time unit.

Breakdown fluency. This dimension can be calculated by observing the number and length of pauses or, vice versa, the time in which a speaker produces language over total time (phonation time ratio). If pauses are counted, then a decision must be made as to how long a silence must be in order to be considered as a pause (typical values are around 0.3-0.4 seconds). Another indicator may be the number of words produced between one pause and another, referred to as "words per burst", "mean length of run" or "mean length of utterance".

The position of pauses is important, too. It has been noted, for example, that native speakers and advanced learners tend to pause between syntactic constituents, mainly to meet conceptual planning needs, while lower-level learners also pause within these constituents, probably to search for words (De Jong, 2016; Tavakoli, 2011).

Repair fluency. This measure concerns the number of repair phenomena (e.g., self-repetitions, reformulations, hesitations, false starts). It gives an indication of the uncertainty experienced by speakers and of how they may consciously monitor their production.

Most research on fluency has concentrated on oral productions. However, the measures presented above can be applied to writing too, provided one has access to keylogging programs. For instance, Gunnarsson (2012) counted the number of words per burst (i.e. between two relatively long pauses), controlling for participants' L1 typing speed, to avoid confounding linguistic fluency with typing skills. It is also possible to calculate the number of keystrokes per minute, which is equivalent to the phonation time ratio.

In the absence of online data regarding the writing process, assessing fluency in written productions becomes more problematic. A gross indicator may be the number of words/characters per minute, provided that the writing time for each composition is known. If this information is missing, and only the final written product is available, a possible measure may be overall text length.

Most studies discussed fluency in monologic productions. However, some researchers have stressed the importance of considering fluency in dialogues too, which is given not only by the fluency of individual

productions, but also by how speakers exchange turns. Peltonen (2017) for instance showed that more advanced learners tend to produce fewer and shorter inter-turn pauses and use more other-repetitions to ensure coherence in dialogue. Sato (2014) has shown that, both in objective measurements and in subjective ratings, individual and interactional fluency are distinct constructs.

Accuracy

While fluency and complexity describe the interlanguage system as such and in positive terms (what is there), accuracy describes it in relation to another system, namely the L2, and in negative terms (what is missing, i.e. the number of errors). This calls into question the very idea of interlanguage as an autonomous system, whose logic and regularities should be explained in their own right, and results in the "comparative fallacy" (Bley-Vroman, 1983) of describing a language in terms of another (e.g. describing Spanish by listing its differences with English), or describing something in terms of what it is not (e.g. describing the language of a three-year old by listing the structures she has not mastered yet). An "interlanguage approach" (Pallotti, 2017) to teaching and assessment, on the contrary, consists in understanding the strategies underlying learners' productions and reconstructing their interlanguage as a system. While this attitude is certainly not as widespread as that of counting and weighing errors (the "accuracy" perspective), it may be relevant especially for formative assessment, and, after some training, teachers can apply it in ordinary classroom contexts (Pallotti, 2010; 2017).

From a psycholinguistic point of view, accurate productions may be seen as depending on the degree of conscious monitoring, so that accuracy would be inversely related to fluency, but this would only apply to occasional mistakes, or slips, while systematic errors due to the interlanguage as such could not be corrected by paying greater attention. Therefore, accurate productions may depend on a more proceduralized norm-oriented interlanguage system, and be thus positively related to increased fluency. Furthermore, accuracy must be assessed by referring to a system of "correct" norms, which is not always easy, due to sociolinguistic variation across social and geographical contexts.

Despite these theoretical difficulties, in practice accuracy is often considered to be an important aspect for language tests' end-users (especially in summative evaluation), as well as being frequently assessed in SLA research.

Accuracy may be operationalized through global measures, such as the average number of errors per unit (clause, sentence, T-Unit, AS-Unit) or, conversely, by the proportion of error-free units. It remains difficult, however, to define exactly what errors are, especially if lexical, morphological, syntactic, phonological or spelling errors are bundled together; moreover, this index is influenced by the average length of units. To overcome this problem, Foster and Skehan (2012) proposed the measure "maximum length of accurate clause", which consists in listing all the clauses in order of length, establishing a cut-off of accuracy (e.g. 70%) and calculating the maximum length of clauses reaching that accuracy threshold.

Other authors counted grammatical, lexical, spelling, appropriateness errors separately (e.g., Kuiken & Vedder, 2012), or proposed more specific measures, such as the proportion of error-free relative clauses, or verbs, or past-tense verbs (Kormos, 2014).

Some have also weighted errors according to their impact on comprehensibility (Foster & Wigglesworth, 2016; Kuiken & Vedder, 2008), or to their concerning more or less complex or advanced structures (Gunnarsson, 2012). However, these measures mix constructs, such as accuracy, on the one hand, and comprehensibility or interlanguage development, on the other, which are logically and, to some extent, empirically different.

The relevance of CAF for language testing

Most tests nowadays are not limited to assessing language proficiency in and for itself, but strive to see it in the context of a larger communicative competence. This points to one of the main limitations of CAF research, which in most cases describes the formal characteristics of linguistic productions with little or no concern for their communicative adequacy. One of the first to raise the issue was Pallotti (2009, p. 596), for whom "adequacy can be seen both as a separate dimension, theoretically independent from CAF (although it may be empirically found to be related to it), and as a way of interpreting CAF measures themselves."

Some studies have followed the first direction, correlating functional adequacy assessments with CAF measures (e.g., de Jong et al., 2012; Révész, Ekiert & Torgersen, 2016). However, the second path is

probably more fruitful for language assessment. Given that the basic construct of most contemporary tests is linguistic-communicative competence, the measurement of Complexity, Accuracy and Fluency should be interpreted according to this fundamental dimension. Thus, a communicatively oriented test should not only assess whether a production is more or less complex, or fluent, but whether and to what extent such complexity and fluency are appropriate for performing communicative tasks. For instance, there may be target domains where complex linguistic structures (e.g. rare words or long sentences) may be unnecessary, or even harmful, so that the more is not necessarily the better.

As regards fluency, Fulcher (2015) argues that most studies that assessed it with measures such as speech rate or pause number and length under-represented the construct, limiting themselves to a purely objective account, with low-inference descriptive categories. If the latter were interpreted at all, this was done in terms of underlying cognitive processes, such as "cognitive fluency" (Segalowitz, 2010). This approach thus neglected the communicative function of pauses, disfluencies and articulation speed, and their role in the co-construction of meaning among participants, which, in his opinion, are the key aspects in a test, and that only a human rater can adequately assess.

Recommendations for practice

Mind your constructs

Fulcher's criticism raises the fundamental questions: what do we want to measure, and how? Terms like *complex*, *accurate* and *fluent* are regularly used in descriptor scales, but many of them are polysemic, vague and with different meanings in the ordinary language versus technical domains. Therefore, a first contribution of CAF research may be to help test developers and users to be more explicit about these constructs. In practice, many rating scales and test scores will be based on a number of these aspects at the same time, but this should be the result of an intended choice, with a clear awareness of what (sub)dimensions have been bundled together and why.

With regard to **complexity**, one needs to specify whether it regards the lexicon, morphology or syntax, or all of them. The complexity of each of these levels should be interpreted in different ways: lexical complexity may index the ability to express precise semantic nuances, morphological complexity concerns the command of a wide range of grammatical structures, while syntactic complexity may have to do with grammatical proficiency, too, but also with cultural and individual rhetorical choices - written academic language tends to favor more syntactically complex texts than professional or mundane usages, but even the ideal or standard level of syntactic complexity in academic texts is not the same in different languages; furthermore, there is always an unavoidable degree of individual stylistic preferences. One should also bear in mind the distinction between complexity and difficulty.

Fluency is a multidimensional construct, too, since it results from different aspects such as speech rate, pauses, reformulations, and because it can be interpreted in different ways. Observable phenomena may be seen as indicating increasingly automatic cognitive processes (cognitive fluency), or as more or less functional behaviors for achieving communicative purposes (perceived fluency). Also, several aspects of L2 fluency (especially length of pauses and articulation rate) are correlated to L1 speaking style, which requires great care when comparing candidates using the same absolute measures (de Jong, 2018).

In a purely communicative approach to language testing, **accuracy** should not be very relevant, for the main aim should be to assess whether candidates are able to perform certain communicative activities, regardless of how they use articles or conjugate verbs. Most errors at the intermediate and advanced levels do not affect comprehensibility and thus communicative effectiveness, while problems at lower levels are better characterized, rather than in terms of errors, in terms of a basic interlanguage system, which may not be adequate to clearly express some lexical and grammatical meanings. This is perhaps why most accuracy scales classify errors according to their impact on comprehensibility, although this mixes two theoretically and empirically distinct dimensions. A classification of errors based on their appearance in a developmental sequence may be useful for placing learners in different classes or course levels; however, in this case I would recommend carrying out a comprehensive interlanguage analysis, which goes beyond simply counting and classifying errors, but includes a systematic description of learners' strategies.

Objective measures and human ratings

Once the constructs have been clearly defined, the next decision concerns how they should be assessed. Most SLA studies have employed objective measures, whereas in language testing performance is usually assessed through rating scales used by human judges (see Kuiken & Vedder, this volume; Pill & Smart, this volume).

The relationship between objective measures and human ratings can be seen in different ways. It could be argued that they concern different constructs: for example, objective fluency measurements would give an indication of individual cognitive processes, while subjective ratings would assess the social and communicative uses of (dis)fluency phenomena. The question then arises as to whether these different constructs, measured in different ways, are all relevant for the test purposes. On the other hand, one may argue that objective measures and human ratings (should) assess the same constructs, with holistic rating just being a more practical alternative than the meticulous counting of errors, pauses or clauses. In this case, the correlation between objective and subjective evaluation needs to be determined: most studies have found them to range from fair to good, but rarely very strong (e.g. Brown, Iwashita & McNamara, 2005).

The analytic study of CAF may be relevant for raters' training, making them aware of subtle distinctions in the constructs they are evaluating. For instance, when assessing complexity, judges seem to pay more attention to general syntactic complexity features, like average length of syntactic constituents or subordination ratio, than to intra-clausal phenomena (Tonkyn, 2012). As regards fluency, measures like articulation rate and length of pauses correlate rather clearly with human ratings, while filled pauses and repairs seem to be less directly related to human perceptions of fluency, possibly because they may be more related to individual speaking style or to task demands (Kormos & Dénes, 2004; Sato, 2014; Tavakoli, Nakatsuhara & Hunter, 2017).

These results thus show one of the main consequences of CAF research for language assessment. CAF analytic measures are unlikely to be directly used to arrive at a test score, both because calculating them is time-consuming (although automatic scoring is becoming increasingly available for some of them), but also, and more importantly, because human judgment is often necessary to determine the communicative value of these linguistic aspects. Nonetheless, fine-grained CAF analysis may be valuable in various ways to language assessment, for instance to validate constructs and assessment procedures, to develop data-based rating scales or for improving assessors' performance, making them aware of the multidimensionality of many test constructs.

Testing tips

In order to incorporate CAF dimensions into language assessment, the following points should be borne in mind.

- Are Complexity, Accuracy and Fluency really relevant for the test construct? For instance, a task-based test with a strong communicative orientation may include task completion as the main or only criterion and consider the linguistic characteristics of candidates' production as irrelevant.
- If CAF dimensions are deemed to be part of the construct to be evaluated, the following considerations may apply.
    - **Complexity**. How important is complexity in the target domain? Complex constructions seem to be highly valued in academic contexts, but may not be so in others; in any case, complexity cannot grow indefinitely and beyond a certain level it may hamper functional communication.
    - **Fluency**. May be related to individual style and task characteristics, so that mechanistic interpretations in terms of 'the more, the better' are problematic.
    - **Accuracy**. Should not be mixed with communicative adequacy or interlanguage development.
- CAF dimensions (together with a thorough interlanguage analysis) may be more appropriate for formative than for summative assessment.

Recommended readings

- Brown, A., Iwashita, N., & McNamara, T. (2005). An examination of rater orientations and test□taker performance on English□for□academic□purposes speaking tasks. *ETS Research Report Series*, *2005* (1). A comprehensive report comparing holistic ratings and CAF measures in the validation of oral language tasks and assessment procedures.
- de Jong, N. H. (2018) Fluency in Second Language Testing: Insights From Different Disciplines, *Language Assessment Quarterly*, *15*(3), 237-254. An exemplary article showing how research on CAF (in this case, fluency) may be applied in language testing contexts.
- Housen, A., Kuiken, F. & Vedder, I. (Eds.). (2012). *Dimensions of L2 performance and proficiency. Complexity, accuracy and fluency in SLA*. Amsterdam: John Benjamins. An influential collection of papers, covering many theoretical and practical issues related to CAF. Some chapters also specifically address the language testing perspective.

References

Bley-Vroman, R. (1983). The comparative fallacy in interlanguage studies: The case of systematicity, *Language learning*, *33*(1), 1-17. doi:10.1111/j.1467-1770.1983.tb00983.x.

Brezina, V., & Pallotti, G. (2019). Morphological complexity in written L2 texts. *Second Language Research*, 35(1), 99–119. doi:10.1177/0267658316643125.

Brown, A., Iwashita, N., & McNamara, T. (2005). An examination of rater orientations and test□taker performance on English□for□academic□purposes speaking tasks. *ETS Research Report Series*, *2005* (1).

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A. Housen, F. Kuiken, & I.

Vedder (Eds.), *Dimensions of L2 performance and proficiency. Complexity, accuracy and fluency in SLA* (pp. 21–46). Amsterdam: John Benjamins. doi:10.1075/lllt.32.02bul.

De Clercq, B. & Housen, A. (2019). The development of morphological complexity: A cross-linguistic study of L2 French and English. *Second Language Research*, 35(1), 71–97. doi:10.1177/0267658316674506

De Clercq, B., & Housen, A. (2017). A cross□linguistic perspective on syntactic complexity in L2 development: Syntactic elaboration and diversity. *The Modern Language Journal, 101*(2), 315-334. doi:10.1111/modl.12396.

de Jong, N. H. (2016). Predicting pauses in L1 and L2 speech: The effects of utterance boundaries and word frequency. *International Review of Applied Linguistics in Language Teaching*, *54*(2), 113-132. doi:10.1515/iral-2016-9993.

de Jong, N. H. (2018) Fluency in second language testing: Insights from different disciplines, *Language Assessment Quarterly*, *15*(3), 237-254, doi: 10.1080/15434303.2018.1477780

de Jong N. H., Steinel M. P., Florijn A., Schoonen R., & Hulstijn J. H. (2012). The effect of task complexity on functional adequacy, fluency and lexical diversity in speaking performances of native and non-native speakers. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency. Complexity, accuracy and fluency in SLA* (pp. 121-142). Amsterdam: Benjamins.

Ellis, R. (2018). *Reflections on task-based language teaching.* Oxford: Oxford University Press.

Foster, P. & Skehan, P. (2012). Complexity, accuracy, fluency and lexis in task-based performance. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency. Complexity, accuracy and fluency in SLA* (pp. 199-220). Amsterdam: John Benjamins. doi:10.1075/lllt.32.09fos.

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied linguistics*, *21*(3), 354-375. doi:10.1093/applin/21.3.354.

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

Fulcher, G. (2015). *Re-examining language testing: A philosophical and social inquiry*. London: Routledge. doi: 10.4324/9781315695518

Gunnarsson, C. (2012). The development of complexity, accuracy and fluency in the written production of L2 French. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency. Complexity, accuracy and fluency in SLA* (pp. 247-276). Amsterdam: John Benjamins. doi:10.1075/lllt.32.11gun.

Horst M and Collins L (2006) From faible to strong: How does their vocabulary grow? *Canadian Modern Language Review/La Revue canadienne des langues vivantes*, *63*(1): 83-106.doi.org/10.3138/cmlr.63.1.83

Housen, A., Kuiken, F. & Vedder, I. (Eds.). (2012). *Dimensions of L2 performance and proficiency. Complexity, accuracy and fluency in SLA*. Amsterdam: John Benjamins.

Hsieh, C.-N., & Wang, Y. (2019). Speaking proficiency of young language students: A discourse-analytic study. *Language Testing*, *36*(1), 27–50. https://doi.org/10.1177/0265532217734240Hunt, K. W. (1965). Grammatical structures written at three grade levels. *NCTE Research Report* No. 3. Paper presented at the National Council of Teachers of English, Champaign, IL.

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.

Kormos, J. (2014). Differences across modalities of performance. In H. Byrnes & R. Manchón (Eds.), *Task-based language learning–Insights from and for L2 writing* (pp. 193-216). Amsterdam: John Benjamins. doi:10.1075/tblt.7.08kor.

Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System*, *32*(2), 145-164. doi:10.1016/j.system.2004.01.001.

Kuiken, F., & Vedder, I. (2008). Cognitive task complexity and written output in Italian and French as a foreign language. *Journal of Second Language Writing*, *17*(1), 48-60. doi:10.1016/j.jslw.2007.08.003.

Kuiken, F., & Vedder, I. (2012). Syntactic complexity, lexical variation and accuracy as a function of task complexity and proficiency level in L2 writing and speaking. In A. Housen, F. Kuiken & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency. Complexity, accuracy and fluency in SLA* (pp. 143-170). Amsterdam: John Benjamins. doi:10.1075/lllt.32.07kui.

Lambert, C., & Kormos, J. (2014). Complexity, accuracy, and fluency in task-based L2 research: Toward more developmentally based measures of second language acquisition. *Applied Linguistics, 35*(5), 606–614. doi:10.1093/applin/amu047.

Malvern D., Richards B.J., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development:*

Quantification and assessment. Basingstoke: Palgrave Macmillan. doi: 10.1057/9780230511804
McCarthy, P. M., & Jarvis, S. (2013). From intrinsic to extrinsic issues of lexical diversity assessment. In S. Jarvis & M. Daller (Eds.), *Vocabulary knowledge: Human ratings and automated measures* (pp. 45–78). Amsterdam: Benjamins.
Michel, M. (2017). Complexity, accuracy and fluency in L2 production. In S. Loewen & M. Sato (Eds.), *The Routledge handbook of instructed second language acquisition* (pp. 50-68). London: Routledge.
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. doi:10.1093/applin/amp044.
Pallotti, G. (2009). CAF: Defining, refining and differentiating constructs. *Applied Linguistics, 30*(4), 590–601. doi:10.1093/applin/amp045.
Pallotti, G. (2015). A simple view of linguistic complexity. *Second Language Research*, *31*(1), 117–134. doi:10.1177/0267658314536435.
Pallotti, G. (2010). Doing interlanguage analysis in school contexts. In I. Bartning, M. Martin, & I. Vedder (Eds.), *Communicative proficiency and linguistic development* (pp. 159–190). Amsterdam: Eurosla Monographs.
Pallotti, G. (2017). Applying the interlanguage approach to language teaching. *International Review of Applied Linguistics in Language Teaching (IRAL)*, *55*(4), 393–412. https://doi.org/doi.org/10.1515/iral-2017-0145
Peltonen, P. (2017). Temporal fluency and problem-solving in interaction: An exploratory study of fluency resources in L2 dialogue. *System*, *70*, 1–13. doi:10.1016/j.system.2017.08.009.
Plonsky, L., & Kim, Y. (2016). Task-based learner production: A substantive and methodological review. *Annual Review of Applied Linguistics*, *36*, 73–97. https://doi.org/10.1017/S0267190516000015
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. doi:10.1016/j.jslw.2014.09.003.
Révész, A., Ekiert, M., & Torgersen, E. N. (2016). The effects of complexity, accuracy, and fluency on communicative adequacy in oral task performance. *Applied Linguistics*, *37*(6), 828–848. doi:10.1093/applin/amu069
Robinson, P. (2015). The Cognition Hypothesis, second language task demands, and the SSARC model of pedagogic task sequencing. In M. Bygate (Ed.), *Domains and directions in the development of TBLT* (pp. 87–121). Amsterdam: Benjamins.
Sato, M. (2014). Exploring the construct of interactional oral fluency: Second language acquisition and language testing approaches. *System*, *45*, 79–91. doi:10.1016/j.system.2014.05.004.
Segalowitz, N. (2010). *The cognitive bases of second language fluency*. New York, NY: Routledge.
Skehan, P. (1998). *A cognitive approach to language learning*. Oxford: Oxford University Press.
Skehan, P. (2015). Limited attentional capacity and cognition: Two hypotheses regarding second language performance on tasks. In M. Bygate (Ed.), *Domains and directions in the development of TBLT* (pp. 123–155). Amsterdam: Benjamins.
Tavakoli, P., & Skehan, P. (2005). Strategic planning, task structure, and performance testing. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 239-273). Amsterdam: John Benjamins. doi:10.1075/lllt.11.15tav.
Tavakoli, P., Nakatsuhara, F. & Hunter, A. M. (2017). Scoring Validity of the Aptis Speaking test: Investigating Fluency across Tasks and Levels of Proficiency. *ARAGs Research Reports Online, AR-G/2017/7*, 1-56.
Thirakunkovit, S., Rodríguez-Fuentes, R. A., Park, K., & Staples, S. (2019). A corpus-based analysis of grammatical complexity as a measure of international teaching assistants' oral English proficiency. *English for Specific Purposes*, *53*, 74–89. doi.org/10.1016/j.esp.2018.09.002
Tonkyn, A. (2012). Measuring and perceiving changes in oral complexity, accuracy and fluency. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency. Complexity, accuracy and fluency in SLA* (pp. 221-244). Amsterdam: Benjamins. doi:10.1075/lllt.32.10ton.
Towell, R. (2012). Complexity, accuracy and fluency from the perspective of psycholinguistic second language acquisition research. In A. Housen, F. Kuiken & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency. Complexity, accuracy and fluency in SLA* (pp. 47-70). Amsterdam: Benjamins.
Wolfe-Quintero, K., Inagaki, S. & Kim, H. Y. (1998). *Second language development in writing: Measures of fluency, accuracy, & complexity*. Honolulu, HI: University of Hawaii Press.