Michel, M. (2017) Complexity, Accuracy and Fluency (CAF). In Shawn Loewen & Masatoshi Sato *The Routledge Handbook of Instructed Second Language Acquisition*. London: Routledge.

Complexity, accuracy, and fluency

Marije Michel

Lancaster University

BACKGROUND

Measuring the product of second language (L2) performance, i.e., oral or written language, is a crucial aspect of research into instructed second language acquisition (ISLA) and has a long tradition. The earliest attempts to gauge performance in modern SLA research emerged in the 1970s and can be divided into two main strands (see Housen, Kuiken, & Vedder, 2012; Wolfe-Quintero, Kim, & Inagaki, 1998): (a) Based on research into first language (L1) acquisition, where mean length of utterance (MLU) was an established index of development, L2 researchers aimed for an index that would allow measurement of global L2 proficiency in reliable and valid ways and that would permit comparability over different studies and languages (see Larsen-Freeman, 1978); (b) From a pedagogical perspective, more and more classroom-based research into L2 performance started to characterize language use in terms of accuracy on the one hand and fluency on the other hand (Brumfit, 1979). Skehan (1989) added complexity and thereby introduced the triad of complexity, accuracy, and fluency (CAF) as the three fundamental dimensions characterizing L2 usage (Housen & Kuiken, 2009).

To date, the early working definitions of CAF are still used for global proficiency: Complexity refers to the size, elaborateness, richness, and diversity of the L2 performance. Accuracy is a measure for the target-like and error-free use of language. Fluency refers to the smooth, easy and eloquent production of speech with limited numbers of pauses, hesitations or reformulations. In the past two decades a growing body of research into ISLA has used CAF measures as dependent variables to gauge L2 performance manipulated by independent variables such as task complexity and task repetition. To a lesser extent some developmental studies have used CAF to identify change in quasi-experimental studies with pre-/post-test designs while others showcase longitudinal learner trajectories (for recent reviews see Housen & Kuiken, 2009, Housen et al., 2012, and Lambert & Kormos, 2014, on CAF in general; Bulté & Housen, 2012, on complexity; Polio & Shea, 2014, on accuracy; Bosker, Pinget, Quené, Sanders, & de Jong, 2013, on fluency).

With respect to ISLA, Norris and Ortega (2009, p. 557) state that “the primary reason for measuring L2 CAF is to account for how and why language competencies develop for specific learners and target languages, in response to particular tasks, teaching, and other stimuli, and mapped against the details of developmental rate, route, and ultimate outcomes.” CAF dimensions are thought to be able to characterize different levels of L2 performance (Wolfe-Quintero et al., 1998). Furthermore, it is often assumed—although this is not always the case (see Lambert & Kormos, 2014; Pallotti, 2009)—that in comparison to less proficient L2 users or to themselves at earlier stages of development (e.g., before an instructional intervention), more proficient L2 learners (or after an instructional intervention):

(a) use a wider range of and more *complex* grammatical structures and vocabulary;

(b) produce more error-free utterances, i.e., they are more *accurate*; and

(c) speak and/or write more *fluently*, i.e., faster and with fewer instances of silence and repair.

In terms of cognitive processing, greater complexity and accuracy have been associated with a more elaborate and sophisticated L2 knowledge system related to representation and restructuring (or development) of the interlanguage while greater fluency is linked to more control and automatization, i.e., faster access to L2 knowledge (Housen et al., 2012; Skehan, 2009).

Key Concepts

- Complexity: size, elaborateness, richness, and diversity of the learner's linguistic L2 system (Housen & Kuiken, 2009)
- Accuracy: degree of deviancy from a particular norm; deviations are usually characterized as errors (Wolfe-Quintero et al., 1998)
- Fluency: ease, eloquence, and smoothness of speech or writing (Chambers, 1997; Freed, 2000; Koponen & Riggenbach, 2000; Lennon, 1990)

The aim of this chapter is to give an overview of the CAF triad. In the next section, each of the three dimensions is presented with a definition, followed by a review of the challenges faced by current research. A final paragraph discusses ways to measure CAF. The following section will review empirical work that employed CAF and that provided experimental evidence relevant to ISLA. The next section will shed light on future directions in CAF research such as the role of communicative adequacy, the value of CAF when gauging interactive performance, and the use of advanced statistical methods and computer-based tools for CAF measurement. Finally, this chapter discusses the need for, on the one hand, standardization to increase validity, reliability, and generalizability of empirical work using CAF, and on the other hand, the need for (new) measures that are able to characterize the dynamic and organic system of L2 production and development.

CURRENT ISSUES

With growing interest to use CAF as dependent variables to measure effects of manipulations on independent variables such as planning time, researchers have started to investigate the constructs more closely, with questions like: What exactly are we evaluating when measuring complexity? What is the 'best' measure to gauge accuracy? What are components of fluency? How do complexity, accuracy, and fluency and their subcomponents interact? Based on these reflections, the early assumption "that these three characteristics of language progress in tandem" (Wolfe-Quintero et al., 1998, p. 4) has now made room for the acknowledgement that complexity, accuracy and fluency are multi-faceted, multi-layerd and multi-dimensional in nature and that they are interrelated in complex and not necessarily linear ways (Housen et al., 2012; Lambert & Kormos, 2014; Larsen Freeman, 2009; Norris & Ortega, 2009).

Complexity

Complexity is seen as the most controversial dimension of the three CAF constructs (Norris & Ortega, 2009; Pallotti, 2009, 2015). The confusion starts with the fact that complexity applies to different aspects of SLA. There is (a) developmental complexity ("the order in which linguistic structures emerge and are mastered in second (and, possibly, first) language acquisition" Pallotti, 2015, p. 2); (b) cognitive complexity (the subjective difficulty of a language feature, that is, how a learner perceives the difficulty of an item as it is processed and learned); and (c) linguistic complexity (objective complexity which refers to "intrinsic formal or semantic-functional properties of L2 elements (e.g., forms, meanings and form-meaning mappings)" Housen et al., 2012, p. 4). To give an example, learners may perceive the English article system (zero, *a/an*,*th*e) as very difficult, and its mastery might only follow at a later stage of development, while linguistically it could be argued to be fairly simple.

When measuring complexity, the linguistic dimension has most often been applied in CAF research. Linguistic complexity itself is a multidimensional construct. In their meticulous examination of L2 complexity, Bulté and Housen (2012, p. 24) define it as "the number of discrete components that a language feature or a language system consists of, and the number of connections between the different components." They make a basic distinction between lexical complexity and grammatical complexity – a view that is in accordance with the body of empirical CAF studies.

Many scholars have set out to disentangle the different sub-dimensions of lexical complexity and to identify appropriate measures (e.g., Jarvis, 2013; Jarvis & Daller, 2013; Malvern & Richards, 1997; Vermeer, 2000). Most work differentiates lexical diversity (i.e., the size of the lexicon measured by means of, for example, type-token ratio measures), lexical sophistication (i.e., the depth of lexis measured by means of, for example, frequency of rare or academic words), and lexical density (i.e., the amount of information in a text, typically measured by the ratio of lexical words per function words). Bulté and Housen (2012) proposed to add compositionality (i.e., the number of formal and semantic components of lexical items) while Jarvis (2013) identified six sub-components of lexical diversity: rarity, volume, variability, evenness, disparity, and dispersion. As can be imagined, providing an encompassing picture of the lexical complexity of L2 data is a challenging endeavor.

Key Concepts

Components of lexical complexity:

- *Diversity*: size of lexis; gauged by means of type-token ratio based measures
- *Sophistication*: depth of lexis; gauged by means of frequency measures, for example, of words beyond the 1000 most common words
- *Density*: information packaging of lexis; gauged by means of, for example, ratio of lexical words per function words

Components of grammatical complexity at different linguistic levels (among others morphology, syntax,):

- *Length*: short vs. long units; gauged by, for example, number of words per clause
- *Variation*: variety of units; gauged by, for example, number of different morphemes used
- *Interdependence*: relation between units; gauged by, for example, coordinated vs. subordinated clauses

Grammatical complexity, too, has different sub-dimensions. Even though most research has focused on syntactic complexity (sentence, clause, phrase), Bulté and Housen (2012) stress the importance of a morphological (inflectional, derivational) and. At all these levels, one can distinguish less from more complex language in terms of length (e.g., longer sentences), variation (e.g., more frequent use of different types of morphemes), and interdependence (e.g., coordination versus subordination).

For both, lexical and grammatical complexity, the choice of which and how many components to employ and what exact measures to use is non-trivial as it will impact on the findings of empirical work. Norris and Ortega (2009) stress that one should avoid using co-linear measures (e.g., type-token ratio AND Guiraud’s index because they both tap into lexical diversity). Instead, they suggest using measures that gauge different sub-components and that are likely to distinguish between theoretically expected differences in the specific context. For example, to examine developmental changes at the syntactic level they propose measuring coordination (e.g., the number of coordinated phrases, a sign of complexification at initial stages of L2 proficiency), subordination (e.g., number of subordinate clauses, a good indicator of complexification at intermediate L2 levels), and phrase-internal complexification (e.g., length of noun phrases, for higher levels of L2 knowledge or L1 data). However, this development view has recently been challenged by Inoue (2016) who found task effects to be more important than proficiency. Similarly, Lambert and Kormos (2014) argue that these measures are not fine-grained enough and instead propose to analyze different types of subordination, e.g., differentiating nominal subordination from subordination via subject/object relative clauses.

A final word of caution about complexity. There is a general tendency to interpret more complexity as an indicator of better language production, for example, more subordination should indicate higher levels of L2 use. However, as discussed by Pallotti (2009) this view is too simplistic. Firstly, linguistic complexity varies by genre (e.g., small talk vs. argumentative essays) and individual stylistic preferences. To quote Pallotti (2009, p. 597): “Beckett is not Joyce, and this has nothing to do with (in)competence, but with stylistic choices.” Secondly, in a dynamic process like L2 development, linguistic complexity cannot be expected to grow linearly (Lambert & Kormos, 2014; Larsen Freeman, 2009). As such, higher complexity (and also fluency) might indicate higher competence or performance levels, but this is by no means an absolute rule.

Accuracy

Accuracy seems to be the most transparent construct in the CAF triad (Housen & Kuiken, 2009; Pallotti, 2009, Wolfe-Quintero et al., 1998), and it refers to target-like-use of language, i.e., error-free speech or writing, and measures the amount of deviation from the norm. The challenge of measuring accuracy is strongly related to the choice of linguistic norm, e.g., a prescriptive grammatical description of the target language or native speaker usage. Applying a linguistic norm raises various issues, e.g., a prescriptive norm might not be appropriate for spoken language use. Furthermore, raters do not always agree on what is accurate (cf. Polio, 1997;

Kuiken & Vedder, 2014). The fact that the same language (e.g., German) may have several normative standards (e.g., Austrian, German, Swiss) adds another layer to this discussion.

Even if there was agreement regarding the norm, there remains the question of how ‘far away’ a deviation from this chosen norm is. For example, a punctuation error may not be as severe as mixing up word order, omitting an article, or using unusual lexical combinations as demonstrated by a comparison of (1) versus (2).

(1) Honestly I think this is an excellent piece of writing.

(2) Honestly, I think this tremendous writing is.

Valid and reliable measures of accuracy should be able to make this distinction (Polio & Shea, 2014). In this sense, Kuiken and Vedder (2008) distinguished 1st, 2nd, and 3rd degree errors in terms of communicative adequacy. Kuiken and Vedder’s (2008) categorization would classify (1) as a 1st degree minor error but (2) as more severe 2nd degree error hampering understanding, while a 3rd degree error would make the sentence incomprehensible, e.g., “Honest, write good.”. More recently, Foster and Wigglesworth (2016) have proposed a weighted measure for accuracy that assigns clauses a score based on their accuracy. Accordingly, the clauses in example (2) would receive the scores 1.0 for ‘Honestly, I think’ and 0.5 for ‘this tremendous writing is’, for an overall score of 1.5. In this way, L2 production can be evaluated quantitatively by assigning a weighted single accuracy score to total performance. However, weighting errors reliably is not an easy task either (Pallotti, 2009), and as Foster and Wigglesworth (2016, p. 112) state: “Anyone who has worked on assessing accuracy in L2 data will know this only too well; some degree of personal judgment has to be invoked occasionally.”

In empirical work, accuracy has been gauged using holistic scales (e.g., Polio, 1997), global measures (e.g., error-free clauses, number of errors per 100 words) as well as specific measures.

The choice for a specific measure will be based on the language that is expected. For example, when investigating the effect of a teaching unit on past tense, target-like-use of past *-ed* would be the specific measure. Similarly, exploring language elicited by a task focusing on plural vs. singular agents might count agreement errors, while the specific L1-L2 combination could make it an obvious choice to go for gender marking on adjectives (for example, for English learners of Spanish).

Each measure comes with its own advantages and shortcomings. Holistic scales allow a global impression which takes into account the severity of errors; however, such scales often do not clearly distinguish accuracy from other dimensions such as complexity (Polio, 1997). Global measures make it possible to compare accuracy over different languages, populations, and tasks. Yet, they might not be sensitive enough to capture slight differences at higher levels of proficiency or of short-term interventions (Lambert & Kormos, 2014). In contrast, specific measures might be able to reveal small changes in accuracy, although it is difficult to generalize the findings to other contexts. Categorizing errors according to severity allows comparisons across studies, but it includes making strong interpretative choices when defining the categories and assigning an error to a certain degree.

Key Concepts

Measuring Accuracy:

- *Holistic scales* provide a global impression of accuracy; for example, low score for “little knowledge of English vocabulary and word forms; virtually no mastery of sentence construction rules; dominated by errors” (Polio, 1997, p. 137)
- *Global* measures quantify overall accuracy; for example, number of error-free clauses.
- *Specific* measures focus on the specific target of a pedagogic intervention, task, or

language; for example, number of noun-adjective-gender-agreement errors.

- *Degrees of errors* weight the severity of an error; for example, 1st degree: minor mistakes like spelling or omitted articles; 2nd degree: more severe mistakes such as word order; 3rd degree: mistakes that make an utterance nearly incomprehensible, e.g., combination of wrong word choice, word order and omissions (cf. Kuiken & Vedder, 2008; Foster & Wigglesworth, 2016)

To recap, even though accuracy seems to be less controversial than complexity, measuring this dimension of L2 use includes taking important decisions about the norm to choose and the severity of a deviance from this norm. In light of these considerations, Housen et al. (2012) appeal for using the abbreviation *A* not only for accuracy but also for appropriateness and acceptability, which would account for language use in different contexts and genres (e.g., *CU 2night* being appropriate in a text message but not in a formal invitation).

Fluency

In contrast to complexity and accuracy, which may pertain to oral and written L2 performance, fluency is first and foremost a measure of spoken language, even though writing research also uses measures of fluency. Historically and informally, the term fluency has been used to characterize a generally proficient L2 speaker (Chambers, 1997). More recent research adheres to a narrower definition (Lennon, 2000) where the construct is thought to encompass cognitive psychological, performative and perceived aspects of fluency (Freed, 2000; Kormos & Dénes, 2004; Segalowitz, 2000, 2010). In ISLA, a definition by Tavakoli and Skehan (2005) is cited regularly, according to which fluency consists of the three sub-dimensions (1) speed or rate, e.g., number of words per minute; (2) silence or breakdown, e.g., amount, location and duration of (filled) pauses; and (3) repair, e.g., false starts, repetitions and self-corrections. In terms of language processing, speed is associated with control of and access to proceduralized knowledge; breakdown is thought to reflect the planning and conceptualization stages of language production; while repair fluency is seen as an indicator of monitoring processes (Levelt, 1989; Segalowitz, 2000, 2010; Skehan, 2003, 2009; Tavakoli & Skehan, 2005).

Key Concepts

Components of fluency:

- *Speed* or *rate*: measured by, for example, syllables per second
- *Silence* or *breakdown*: measured by, for example, number, duration and location (at clause boundaries vs. mid-clause) of pauses
- *Repair*: measured by, for example, number of false starts, repetitions and self-repairs

Measures of fluency based on temporal aspects of speech are relatively uncontroversial to identify and quantify in empirical research (Chambers, 1997), for example by calculating the ratio of syllables per second or the number of repairs per hundred words. It is important, however, to acknowledge that some aspects of fluency have been found to be trait-like personal characteristics rather than indicators of L2 competence (de Jong et al., 2015). De Jong, Steinel, Florijn, Schoonen, and Hulstijn (2012) advocate the use of phonation time ratio ("the percentage of time spent speaking as a percentage proportion of the time taken to produce the speech sample", Kormos & Dénes, 2004, p. 148) instead of silence measures (see also Bosker et al., 2013 for a recent discussion). Moreover, Kormos and Dénes (2004) investigated the relationship between fluency measures and expert ratings of fluency, which revealed that boundaries between fluency, on the one hand, and complexity, and accuracy, on the other hand, are less clear-cut.

For writing, fluency is a more controversial construct because the reiterative process permits planning, monitoring, and editing (Johnson, Mercado, & Acevedo, 2012; Wolfe-Quintero et al., 1998). Typically, the oral measures of speed and breakdown are substituted by metrics of rate (e.g., number of words per minute based on the final text produced) and length (e.g., number of words per utterance). Yet newer studies employed key-stroke logging software (Leijten & van Waes, 2013; Révész, Kourtali, & Mazgutova, in press) that records online writing features like number of characters typed between pauses or the ratio of number of characters produced during writing over the number of characters in the final text. Such measures make it less difficult to identify and disentangle the sub-dimensions of fluency from accuracy and complexity in written performance because they allow to review the process of writing fluency and not a product only.

To sum up, fluency is also a multi-faceted construct with subcomponents. In particular in L2 writing, fluency constitutes a challenging dimension to measure and to conceptualize.

Measuring CAF

By now it has become clear that choosing measures of complexity, accuracy, and fluency needs careful considerations. Similarly, interpretations of results require caution and awareness of the explanatory power and limitations of the metrics used (Norris & Ortega, 2009). In this chapter, no attempt is made to provide a list of the ‘best’ measures. Instead, some thoughts that guide the choice for or against a specific metric are shared.

Measures of CAF come in a variety of forms. Wolfe-Quintero et al. (1998) identify three types: (a) frequency counts of a specific linguistic unit, e.g., number of word tokens; (b) ratio measures, that divide a specific unit by the total number of another unit, e.g., type/token ratio (TTR); and (c) indices, that are calculations of a score by means of a more complex formula, e.g., D is based on “mathematically modelling how new words are introduced into larger and larger language samples” (Malvern & Richards, 2002: 85).

The choice for a metric type will be based on the L2 data under investigation. For example, raw frequencies (e.g., total number of errors) allow comparisons only of L2 samples that are of equal length (e.g., texts of 300 words exactly). As soon as samples differ in length, ratios or indices should be used. Indices are calculated because some ratios are known to be non-linearly affected by sample length (e.g., D, Malvern & Richards, 1997, 2002; or Measure of Textual Lexical Diversity, MTLD, McCarthy & Jarvis, 2010; both adjust TTR for sample length).

When calculating ratios and indices, an important decision is what unit of reference to use (e.g., sentences, clauses, words, minutes, seconds). While research into writing may count sentences (i.e., text between two period marks) as syntactic units, it is difficult to establish ‘sentence’ boundaries in oral performance. Alternative syntactic units include the terminal (T) unit (Hunt, 1965), the communication (C) unit (similar to T unit but including utterances without a verb; Bardovi-Harlig, 1992) and more recently the analysis of speech (AS) unit (Foster, Tonkyn, & Wigglesworth, 2000). The latter has become the standard for oral data (see also Crookes, 1990, for a discussion of different units).

Key Concepts

- <u>Terminal (T) unit</u>: (Hunt, 1965, p. 735): “one main clause plus whatever subordinate clauses happen to be attached or embedded within it”
- <u>Analysis of Speech (AS) unit</u> (Foster, Tonkyn, & Wigglesworth, 2000, p. 365): “a single speaker’s utterance consisting of an independent clause, or sub-clausal unit, together with any subordinate clause(s) associated with either.”

It is advisable to use to some extent the same measures as key references in earlier research to enable comparisons across studies. However, these should be supplemented by measures that are chosen specifically for the current study guided primarily by the type of data and the research questions. For example, Tonkyn (2012) employed eight specific structural measures because he examined development after a short-term intervention, and global measures may not have revealed a change. Michel (2013) decided to calculate the number of conjunctions per 100 words (and not per syntactic unit) in order to avoid interdependence of measures: conjunctions are used to introduce clauses and therefore correlate with the number of syntactic units. For practicality, de Jong et al. (2012) decided to exclude the location of pauses because they used an automatic script (de Jong & Wempe, 2009) to detect pauses in their analysis of over 2000 speech samples and the script did not provide location information.

To summarize, unless L2 samples are of exact equal length, it is advisable to employ ratios and indices rather than raw frequencies. Denominators for these ratios will differ across different measures. Grammatical complexity and accuracy are traditionally expressed as a ratio per syntactic unit (e.g., errors per AS unit). Lexical measures typically take as denominator the number of words (tokens), while fluency employs temporal units such as minutes.

EMPIRICAL EVIDENCE

CAF measures have been used to examine L2 performance, proficiency, and development in a wide variety of fields, including work investigating learner internal factors such as personality (e.g., DeWaele & Furnham, 2000) and age (e.g., Munoz, 2006), as well as external factors such as a specific instructional interventions (e.g., Derwing & Rossiter, 2003; Tavakoli, Campbell, & McCormack, 2015), the learning context (Housen et al., 2011; Mora & Valls-Ferrer, 2012) and many others. This section presents a selective review of empirical work with a focus on studies into different task design and task condition factors that can be manipulated in the classroom. Finally, some work that has used a longitudinal design is presented to provide a developmental perspective.

Task complexity

Task complexity, i.e., the cognitive demands of a task, has received ample attention over the past two decades, particularly, in empirical research investigating the claims of Robinson's (2001) Cognition Hypothesis and Skehan's (1998) Limited Attentional Capacity Model. In short, Skehan predicts that higher cognitive task demands will inevitably result in trade-off effects, in particular between complexity and accuracy, due to competition for limited attentional resources (see Skehan, 2009, for the rationale based on Levelt, 1989). On the contrary, Robinson claims that parallel increases of complexity and accuracy are possible under certain conditions of task design (e.g., when a task requires more reasoning) because the higher cognitive demands require more focused linguistic performance.

Over the years, many studies have set out to contribute to the debate (e.g., the studies gathered in Robinson, 2011). Yet, as Jackson and Suethanapornkul's (2013) research synthesis shows, no compelling answers have been found due, in part, to the large variety of research designs and a plethora of CAF measures generating conflicting results. A meta-analysis of nine comparable studies (Jackson & Suethanapornkul, 2013) revealed that an increase of task complexity resulted in small positive effects for accuracy and small negative effects for fluency (a finding that is consistent with both hypotheses, cf., Skehan, 2009) while grammatical complexity was affected negatively and lexis positively. The latter two findings, however, were not robust enough to support or reject either of the two claims. By synthesizing the findings of seven of their earlier investigations, Skehan and Foster (2012) come to a similar conclusion, i.e., they cannot present firm generalizations because the variety of instruments and measures offered different information.

It is in light of these inconclusive findings from numerous studies that Long (2016) reiterates Norris and Ortega's (2009) call for more standardization and a unified approach to the investigation of task complexity in the future.

Teaching Tips

- It is important to be aware that L2 users are likely to be less fluent when confronted with more complex tasks. However, the higher cognitive demands are likely to result in more accurate and/or complex language and instructors and learners can monitor these to evaluate progress.
- Task repetition and familiarity is a fruitful way to foster higher levels of performance in terms of CAF. Repeating a task just once may enhance their fluency. If targeting accuracy and complexity, multiple task repetitions might be needed to let students overcome trade-off effects between these two dimensions.
- Planning time can be given before (strategic pre-task) or during (unpressured within-task) performance. Giving pre-task planning time is likely to increase complexity and fluency because L2 speakers can conceptualize their performance beforehand. Giving students time to perform a task at their own pace (within-task planning time) decreases fluency but will positively affect complexity and/or accuracy (presumably not both due to trade-off effects).

Task repetition and familiarity

More systematicity in experimental design might be found in the body of research (e.g., Ahmadian & Tavakoli, 2011; Bygate 1996, 2001; Kim & Tracy-Ventura, 2013; Mackey, Kanganas, & Oliver, 2007; Pinter, 2005) that looked into effects of task familiarity and task repetition, i.e., “repetitions of the same or slightly altered tasks – whether whole tasks, or parts of a task” (Bygate & Samuda, 2005, p. 43). Many of these investigations employed CAF measures to evaluate L2 performance. Accordingly, when adults and young learners performed the same or a familiar task more than once, they were more fluent. Findings for complexity and accuracy have resulted in less clear patterns. As Bygate and Samuda (2005) hypothesize, repeated encounters allow L2 performers to shift from meaning-oriented towards more form-oriented production, the latter potentially creating trade-off effects between linguistic complexity and accuracy (Skehan 2009). Overall, though, students’ performance seems to improve when they work more than once on the same or similar material and CAF scores increase accordingly. Using slightly different content for similar tasks (i.e., task familiarity) seems to sustain students’ motivation and interest over multiple repetitions.

Planning time studies

Also, providing L2 users with planning time seems to lead to higher levels of performance, in particular with respect to fluency. Effects of planning time (pre-task planning, within-task planning, task rehearsal) on CAF has been extensively investigated and includes work into oral as well as written production (e.g., Ellis & Yuan, 2005; Foster & Skehan, 1999; Ortega, 2005). Mehnert (1998) showed that an absence of planning-time resulted in low fluency scores, while different lengths of pre-task planning time seemed not to make a difference. In his introduction to an edited volume on planning, Ellis (2005) summarizes that strategic pre-task planning positively affects complexity and fluency while effects on accuracy are mixed. Skehan and

Foster's (2012) synthesis of their earlier work indicates that pre-task planning time affects the conceptualizing stage of speech performance and, therefore, promotes mainly structural complexity and lexical sophistication but also accuracy. The various aspects of fluency (speed, pauses, repair) were found to be affected in different ways.

Recently, Hsu (2015) looked into the effects of planning time in written synchronous computer-mediated communication (SCMC or text chat). Pre-task rehearsal planning time was operationalized as writing a picture description during 10 minutes immediately before 'telling' that story to an SCMC interlocutor. Results showed that rehearsal planning time increased accuracy while complexity seemed to be unaffected. Regarding unpressured, within-task planning, the studies gathered in Ellis (2005) indicate that it promotes accuracy and also complexity while fluency decreases, a finding that was recently replicated by Ahmadian (2012).

To summarize, planning time seems to support conceptualizing (pre-task) and monitoring (within-task), which has the potential to lead to higher scores on all three CAF dimensions. However, trade-off effects are likely to become visible, e.g., increased accuracy as a result of monitoring during within-task planning time might come at the cost of fluency.

Modality: CAF in oral vs. written vs. computer mediated communication

In contrast to the large amount of work on planning time, there are only a handful of CAF studies that have explored effects of different modalities (oral, written, computer-mediated) on L2 performance. Using a between-participant design, Kuiken and Vedder (2012) compared oral versus written production at different levels of task complexity. Their results showed only minor differences between the two modalities. Ellis and Yuan (2005) looked at effects of planning conditions in oral versus written performances. They found greater complexity and accuracy but lower fluency in writing, which they attributed to the fact that writing allows for more planning, formulating, executing and monitoring than speaking.

Sauro (2012) compared oral and written SCMC interactions of L2 speakers. Using measures of complexity and accuracy, no significant differences between the two modes could be attested in group comparisons. Yet, in individual evaluations, large variation between participants emerged, which Sauro assigned to discourse style and turn-taking behaviour as well as typing skills.

In sum, these studies seem to suggest that CAF is not so much affected by modality apart from the expected effects of increased planning time and monitoring during writing.

Teaching Tips

- Be aware that ‘more’ (e.g., complex, fluent) does not automatically entail ‘better’.
- In addition to CAF, there are good reasons to measure performance in terms of communicative adequacy and task completion.
- Some aspects of language use have shown to be related to individual characteristics of a speaker (e.g., syllable duration) and/or are elicited by a specific genre or task feature (clause length). Therefore, such features may not be suitable indicators of proficiency.
- Lack of improvement on one dimension does not mean there is no improvement. Many studies suggest trade-off effects between complexity, accuracy, and fluency (de Jong, Groenhout, Schoonen, & Hulstijn, 2015; Pallotti, 2009).

Longitudinal development

The majority of developmental CAF research has looked into the three dimensions using cross-sectional designs, with just a few recent longitudinal studies.

In writing research, Spoelman and Verspoor (2010) used analytical tools from dynamic systems theory (DST: e.g., Monte Carlo simulations) to explore 54 writing samples of a single learner studying Finnish during three years. Although complexity and accuracy of Finnish case marking showed growth over time, development was non-linear. That is, the data revealed peaks, regressions, and backsliding on specific dimensions, as well as complex interactional patterns among the three dimensions. In another study, Gunnarsson (2012) followed the development of CAF in the written performance of five Swedish L2 learners of French over a period of 30 months. Analyses revealed large individual differences pointing to trade-off effects (Skehan, 2009), i.e., while some writers showed gains in accuracy at the expense of fluency, others prioritized fluency at the cost of accuracy. Polio and Shea (2014) focused on the development of accuracy in a corpus of ESL learners who received writing instruction over the course of one semester (Polio, 1997). They found minor improvements of accuracy but increased linguistic complexity, which they interpret as a trade-off effect. The corpus-based study by Vyatkina, Hirschmann, and Golcher (2015) used multilevel modeling to investigate syntactic development of seven different modifiers (e.g., adverbs, prepositional phrases) in longitudinal writing data of English learners of German over the course of four semesters. This study showed that the global use of modifiers remained relatively stable but the type of modification revealed large inter- and intra-individual variation over time.

Investigations into oral performance include Ferrari (2012), who looked into the development of CAF in four adolescent L2 learners and two native speakers of Italian who performed monologic and dialogic tasks over the course of three years. Her findings suggested trade-off effects between different CAF components in different communicative situations but generally, monologic tasks created greater complexity but lower fluency than dialogic performances. Based on a detailed comparison of the L2 and L1 data, Ferrari concluded that "the ability to vary one's language according to the demands of different communicative activities develops very slowly" (p. 294). In contrast, Vercellotti (2015) could not detect trade-off effects in her data on the oral performance of 66 L2 learners who were recorded monthly over a period of ten months during an intensive English program. Using hierarchical linear modeling she found that grammatical complexity, accuracy, and fluency showed steady linear growth while lexical variety revealed a non-linear trajectory, i.e., there was a dip followed by a steep increase. The case study of, Polat and Kim (2014) who interviewed one uninstructed L2 speaker biweekly during a full year and used dynamic systems theory methods to gain insights into complexity and accuracy development. While lexical complexity showed steady growth over time and syntactic complexity somewhat increased, accuracy seemed unaffected.

FUTURE DIRECTIONS

The role of communicative adequacy

Even though research into CAF suggests that the triad appropriately captures relevant aspects of L2 performance, a call for the inclusion of communicative or functional adequacy has been issued more than once in recent years. Pallotti (2009, p. 596) defines this fourth construct as "the degree to which a learners' performance is more or less successful in achieving the task's goals efficiently." For instance, an utterance scoring high on all three CAF measures can be communicatively inadequate and vice versa, which shows the independence of the two constructs. In language pedagogy and testing, communicative adequacy is one of the main goals, as evidenced for example by the Common European Framework of Reference (CEFR).

To date, only a handful of studies have looked at CAF and communicative adequacy, revealing that they are complementary constructs interacting in several ways. Kuiken, Vedder, and Gilabert (2010) showed that adequacy ratings on L2 writing were not so much correlated to structural complexity, while lexical complexity and accuracy were. Révész, Ekiert, and Togerson (2014) employed linear mixed effects regression and Rasch analyses to investigate adequacy in spoken performance. In their data, the number of filled pauses (i.e., breakdown fluency) seemed to be the strongest predictor of communicative adequacy, while other CAF measures showed minor effects only. Yet another study (de Jong et al.,2012) identified vocabulary knowledge and correct sentence intonation as the strongest predictors of adequacy by means of structural equation modeling,.

CAF in interaction

A disregarded issue in past research has been how the CAF triad accounts for differences between dialogic and monologic performance (but see Ferrari, 2012, reviewed above). Among the few studies, Michel, Kuiken, and Vedder (2007) and Michel (2011) gave the same tasks to L2 (and L1) speakers of Dutch working either on their own or in pairs. Dialogic performance in both populations was characterized by lower grammatical complexity, but higher accuracy and fluency. While non-native speakers were lexically more varied, native speakers showed lower lexical variety in dialogues. Similarly, Gilabert, Barón, and Levkina (2011) found dialogic performances to be more fluent but grammatically less complex.

From a methodological perspective, these studies raise the question of whether current CAF measures gauge the same constructs in monologues and dialogues and whether measurement is valid and reliable. Indeed, both Sato (2014) and Tavakoli (2016), who focus on fluency, state that we might need other measures in dialogues that account for interactive turn-taking patterns because fluency in individual versus interactional performance is fundamentally different. Tavakoli compared several established and newly developed measures of fluency when evaluating monologic and dialogic L2 speech. Findings showed that well-known fluency metrics for monologic production (e.g., phonation time ratio) might not be reliable measures in dialogue, because overlapping speech and between-speaker pauses need to be divided over partners. Earlier, Sato (2014) had already established that raters take into account effective scaffolding and disruptive pause behavior in dialogic speech when assigning fluency scores to speech samples.

Measuring instructional effects by means of CAF

To date surprisingly few studies have used CAF to gauge instructional effects. One reason could be the earlier mentioned concern that global CAF measures might not be sensitive enough to capture slight differences of performance after (short-term) pedagogical interventions. Another cause could be the fact that many interventions focus on a specific linguistic target and, therefore, structure-focused pre-/post tests –rather than global CAF performance measures– are thought to be more suitable. Exceptions are the above mentioned work by Mora and Valls-Ferrer (2012), Tavakoli, Campbell and Cormack (2015), as well as Tonkyn (2012) and other chapters in the edited volumes by Housen et al. (2012) and Baralt et al. (2014). With the development of more fine-grained measures (for example, the ones proposed by Lambert and Kormos, 2014, for syntactic and by Jarvis, 2013, for lexical complexity, respectively) and scores (for example, Foster & Wigglesworth', 2016, weighted accuracy score) future work will hopefully aim to capture instructional effects by means of CAF. The use of CAF measures in future ISLA studies might be further promoted by the growing availability of computerized tools that provide fast and reliable ways to measure CAF. The next section will highlight a few of these tools, knowing the risk of obsolescence due to rapid developments in this area.

Computer-based tools and corpus-based techniques for analyzing CAF

For syntactic complexity, Coh-Metrix (McNamara, Louwerse, Cai, & Graesser, 2013) and Synlex (Lu, 2010), which produce output metrics for length of syntactic units as well as coordination, subordination and syntactic sophistication in L2 writing are widely used. Many (web-based) tools exist that provide calculations of type/token ratios and other measures of lexical diversity, sophistication and density (among others AntWordProfiler, Anthony, 2015; LexTutor for English and French, Cobb, 2000). Fortunately, language corpora are often error-tagged which allows automatic accuracy measurement. However, automatic computer-based accuracy measurement remains a desideratum.

Teaching Tip

- Let students use software (e.g., Synlex and LexProfiler) to analyze the changing complexity of their writing, for example, over tasks, genre and time. Exploring complexity is likely to raise their awareness that accuracy is only one aspect of L2 performance. That is, it might help them to realize that they are making progress in terms of complexity even though error rates do not suggest development.

For fluency, CLAN (MacWhinney, 2000) and Praat (Boersma & Weening, 2013) are widely used. The language independent Praat-script that counts the number of syllables and silent pauses (de Jong & Wempe, 2009) is particularly relevant for the fluency analysis of oral data. Key stroke logging programs such as InputLog (Leijten & van Waes, 2013; see also www.writingpro.eu) allow the investigation of speed, pause and revision measures in computer-based writing, which promises to boost future work into L2 writing by means of CAF.

Finally, corpus-based research facilitates the analysis of developmental trajectories based on large amounts of data (Alexopoulou et al., submitted; Thewissen, 2013; Vyatkina et al., 2015). As more corpora and tools for different languages become available, computer-based CAF research faces a promising future. In the same vein, it is hoped that future longitudinal research will be able to uncover further developmental patterns and individual trajectories of CAF in oral and written L2 performance. From a pedagogic perspective, more future work is needed into the complex interrelationship between communicative adequacy and CAF, in particular, in dialogic settings given that L2 instruction often involves pair work.,

CONCLUSION

Researchers seem to agree that the CAF triad is a useful and valid way to investigate and describe L2 performance and development. However, to date, no consensus has been reached on how to define and measure the constructs.

Over the past decades, many have set out to identify ‘the best’ or ‘a better’ measure (e.g., Kormos & Dénes, 2004; Pallotti, 2015; Polio, 1997; Wolfe-Quintero, 1998). Although these investigations add to our knowledge and understanding, a result is that there are a daunting number of metrics available. For example, Long (2015) criticizes the fact that 84 different measures have been used to examine effects of task complexity. In addition, little is known about the validity and reliability of many measures because most research has paid little attention to these issues. As outcomes are based on different metrics of unknown reliability and validity, it is difficult to identify general trends and compare findings. Consequently, the future calls, on the one hand, for greater standardization and theory-driven use of constructs and metrics and, on the other hand, for the acknowledgement of variability and dynamicity of CAF in L2 language use (Housen et al., 2012; Norris & Ortega, 2009).

REFERENCES

Ahmadian, M. J. (2012). The effects of guided careful online planning on complexity, accuracy and fluency in intermediate EFL learners' oral production: The case of English articles. *Language Teaching Research, 16*(1), 129–149.

Ahmadian, M. J., & Tavakoli, M. (2011). The effects of simultaneous use of careful online planning and task repetition on accuracy, complexity, and fluency in EFL learners' oral production. *Language Teaching Research, 15*(1), 35–59.

Anthony, L. (2015). AntWordProfiler (Version 1.4.1) [Computer program]. Tokyo, Japan: Waseda University. http://www.laurenceanthony.net/

Alexopoulou, T., Meurers, D., Michel, M., & Murakami, A. (submitted) Analyzing learner language in task contexts: A study case of task-based performance in EFCAMDAT.

Baralt, M., Gilabert, R., & Robinson, P. (Eds.) (2014). Task sequencing and instructed second language learning. London: Bloomsbury.Bardovi-Harlig, K. (1992). A second look at T-unit analysis: Reconsidering the sentence. *TESOL Quarterly, 26*(2), 390-395.

Bardovi-Harlig, K. (1992). A second look at T-unit analysis: Reconsidering the sentence. *TESOL Quarterly, 26(2)*, 390-395.

Boersma, P. & Weenink, D. (2013). *Praat: doing phonetics by computer* [Computer program]. http://www.praat.org/

Bosker, H., Pinget, A., Quené, H., Sanders, T. & de Jong, N. (2013). What makes speech sound fluent? The contributions of pauses, speed and repairs. *Language Testing*, *30*, 159–75.

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

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. 23-46). Amsterdam/Philadelphia: John Benjamins.

Bulté, B., & Housen, A. (2014). Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of Second Language Writing, 26,* 42-65.

Bygate, M. (1996). Effects of task repetition: appraising the developing language of learners. In Willis, J., & Willis, D. (Eds.), *Challenge and change in language teaching* (pp. 136-146). Oxford: MacMillan Heinemann.

Bygate, M. (2001). Effects of task repetition on the structure and control of oral language. In van den Branden, K., Bygate, M., & Norris, J. (Eds.), *Task-based language teaching: A reader. (Vol. I*, pp. 249-274). Amsterdam/Philadelphia: John Benjamins.

Bygate, M. & Samuda, V. (2005). Integrative planning through the use of task-repetition. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 37-74). Amsterdam/Philadelphia: John Benjamins.

Chambers, F. (1997). What do we mean by fluency? *System, 25*(4), 535-544.

Cobb, T. Web Vocabprofile, an adaptation of Heatley, Nation & Coxhead's (2002). Range. [Computer program]. http://www.lextutor.ca/vp/

Crookes, D. (1990). The utterance and other basic units for second language discourse analysis. *Applied Linguistics 11*, 183-199.

Dewaele, J. M., & Furnham, A. (2000). Personality and speech production: a pilot study of second language learners. *Personality and Individual Differences, 28*(2), 355-365.

Derwing, T. M., & Rossiter, M. J. (2003). The effects of pronunciation instruction on the accuracy, fluency, and complexity of L2 accented speech. *Applied Language Learning, 13(1),* 1-17.

Ellis, R. (Ed.). (2005). *Planning and task performance in a second language*. Amsterdam/Philadelphia: John Benjamins.

Ellis, R. & Yuan, F. (2005). The effects of careful within-task planning on oral and written task performance. In R. Ellis (Ed.), *Planning and task-based performance in a second language (pp. 167-192).* Amsterdam/Philadelphia: John Benjamins.

Ferrari, S. (2012). A longitudinal study of complexity, accuracy and fluency variation in second language development. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency* (pp. 277-298). Amsterdam/Philadelphia: John Benjamins.

Foster, P. & Skehan, P. (1999). The effect of source of planning and focus on planning on task-based performance. *Language Teaching Research, 3*(3), 185–215.

Foster, P., Tonkyn, A. & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics, 21*(3), 354–375.

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.

Freed, B. (2000). Is fluency, like beauty, in the eyes (and ears) of the beholder. In H. Riggenbach (Ed.), *Perspectives on fluency (*pp. 243-265*).* Michigan: University of Michigan Press.

Gilabert, R., Barón, J. & Levkina, M. (2011). Manipulating task complexity across task types and modes. In P. Robinson (Ed.), *Second language task complexity. Researching the Cognition Hypothesis of language learning and performance* (pp. 105–138). Amsterdam/Philadelphia: John Benjamins.

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* (pp. 247–276). Amsterdam/Philadelphia: John Benjamins.

Housen, A., & Kuiken, F. (2009). Complexity, accuracy and fluency in second language acquisition. *Applied Linguistics, 30*(4), 461-473.

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

Housen, A., Schoonjans, E., Janssens, S., Welcomme, A., Schoonheere, E., & Pierrard, M. (2011). Conceptualizing and measuring the impact of contextual factors in instructed SLA–the role of language prominence. *IRAL-International Review of Applied Linguistics, 49*(2), 83-112.

Hunt, K. W. (1965). *Grammatical structure written at three grade levels* (Research Report 3). Urbana, IL: National Council of Teachers of English.

Hsu, H. C. (2015). The effect of task planning on L2 performance and L2 development in text-based synchronous computer-mediated communication. *Applied Linguistics*, 1-28.

Inoue, C. (2016). A comparative study of the variables used to measure syntactic complexity and accuracy in task-based research. *The Language Learning Journal*, 1-19.

Jackson, D. O., & Suethanapornkul, S. (2013). The Cognition Hypothesis: A synthesis and meta-analysis of research on second language task complexity. *Language Learning, 63*(2), 330–367.

Jarvis, S. (2013). Capturing the diversity in lexical diversity. *Language Learning, 63*(s1), 87-106.

Jarvis, S., & Daller, M. (Eds.). (2013). Vocabulary knowledge: Human ratings and automated measures. Amsterdam/Philadelphia: John Benjamins.

Johnson, M. D., Mercado, L. & Acevedo, A. (2012).The effect of planning sub-processes on L2 writing fluency, grammatical complexity, and lexical complexity. *Journal of Second Language Writing, 21*, 264-282.

de Jong, N. H., Groenhout, R., Schoonen, R., & Hulstijn, J. H. (2015). Second language fluency: Speaking style or proficiency? Correcting measures of second language fluency for first language behavior. *Applied Psycholinguistics*, *36,* 223-243.

de Jong, N. H., Steinel, M., Florijn, A., Schoonen, R., & Hulstijn, J. (2012). Facets of speaking proficiency*, Studies in Second Language Acquisition*, *34*(1), 5–34.

de Jong, N. H. & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods, 41*(2), 385-390.

Kim, Y.J., & Tracy-Ventura, N. (2013). The role of task repetition in L2 performance development: What needs to be repeated during task-based interaction? *System*, *41*(3), 829–840.

Koponen, M., & Riggenbach, H., (2000). Overview: Varying perspectives on fluency. In H. Riggenbach (Ed.), *Perspectives on fluency (*pp. 5–24*).* Michigan: University of Michigan Press.

Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System*, *32, 145–64.*

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.

Kuiken, V., & 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. 143-169). Amsterdam: John Benjamins.

Kuiken, F., & Vedder, I. (2014). Rating written performance: What do raters do and why? *Language Testing, 31*(3), 329-348.

Kuiken, F., Vedder, I., & Gilabert, R. (2010). Communicative adequacy and linguistic complexity in L2 writing. In I. Bartning, M. Martin, & I. Vedder (Eds), *Communicative proficiency and linguistic development: Intersections between SLA and language testing research* (pp. 81-100). Eurosla Monographs 1. Eurosla.

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), 607–614.

Larsen Freeman, D. (1978). An ESL index of development. *TESOL Quarterly, 12*(4), 439-448.

Larsen Freeman, D. (2009). The emergence of complexity, fluency, and accuracy in the oral and written production of five Chinese learners of English. *Applied Linguistics*, *30*(4), 590-619.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation.* Cambridge, MA: MIT Press.

Leijten, M., & Van Waes, L. (2013). Keystroke logging in writing research: Using Inputlog to analyze and visualize writing processes. *Written Communication, 30*(3), 358–392.

Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language learning, 40*(3), 387-417.

Long, M. H. (2015). Second language acquisition and task-based language teaching. Oxford, UK: Wiley-Blackwell.

Long, M. H. (2016). In defense of tasks and TBLT: Nonissues and real issues. *Annual Review of Applied Linguistics, 36,* 5–33.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics, 15*(4), 474-496.

Mackey, A., Kanganas, A., & Oliver, R. (2007). Task familiarity and interactional feedback in child ESL classrooms. *TESOL Quarterly, 41*(2), 285-312.

MacWhinney, B. (2000). The childes project: Tools for analyzing talk. [Computer Program]. http://childes.psy.cmu.edu/

Malvern, D., & Richards, B. (1997). A new measure of lexical diversity. In A. Ryan & A. Wray (Eds.), *Evolving models of language* (pp. 58–71). Clevedon, UK: Multilingual Matters.

Malvern, D., & Richards, B. (2002). Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. *Language Testing, 19(1)*, 85-104.

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods, 42*(2), 381-392.

McNamara, D. S., Louwerse, M. M., Cai, Z., & Graesser, A. (2013). Coh-Metrix version 3.0. http://cohmetrix.com

Mehnert, U. (1998). The effects of different lengths of time for planning on second language performance. *Studies in Second Language Acquisition, 20*(1), 52–83.

Michel, M., Kuiken, F., & Vedder, I. (2007). The influence of complexity in monologic versus dialogic tasks in Dutch L2. *International Review of Applied Linguistics in Language Teaching, 45*(3), 241-259.

Michel, M. (2011) Effects of task complexity and interaction on L2 performance. In P. Robinson (Ed.), *Second language task complexity: Researching the cognition hypothesis of language learning and performance* (pp. 141-174). Amsterdam/Philadelphia: John Benjamins.

Michel, M. (2013). Effects of task complexity on the use of conjunctions in oral L2 task performance. *The Modern Language Journal*, *97*(1), 178–195.

Mora, J. C., & Valls-Ferrer, M. (2012). Oral fluency, accuracy, and complexity in formal instruction and study abroad learning contexts. *TESOL Quarterly*, *46*(4), 610-641.

Munoz, C. (Ed.). (2006). *Age and the rate of foreign language learning*. Multilingual Matters.

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.

Ortega, L. (2005). What do learners plan? Learner-driven attention to form during pre-task planning. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 77-109). Amsterdam/Philadelphia: John Benjamins.

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

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

Pinter, A. (2005). Task repetition with 10-year-old children. In C. Edwards & J. Willis (Eds.), *Teachers exploring tasks in English language teaching* (pp. 113-126). Basingstoke: Palgrave Macmillan.

Polat, B., & Kim, Y. J. (2014). Dynamics of complexity and accuracy: A longitudinal case study of advanced untutored development. *Applied Linguistics*, *35*(2), 184-207.

Polio, C. G. (1997). Measures of linguistic accuracy in second language writing research. *Language Learning, 47*, 101–143.

Polio, C. G., & 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.

Révész, A., Ekiert, M., & Torgersen, E. (2014). The effects of complexity, accuracy, and fluency on communicative adequacy in oral task performance. *Applied Linguistics,* 1-22.

Révész, A., Kourtali, N. & Mazgutova, D. (in press). The effects of task complexity on L2 writing processes, behaviours and linguistic complexity. *Language Learning.*

Robinson, P. (2001). Task complexity, task difficulty and task production: Exploring interactions in a componential framework. *Applied Linguistics, 22*(1), 27–57.

Robinson, P. (Ed.). (2011). Second language task complexity: Researching the cognition hypothesis of language learning and performance. Amsterdam/Philadelphia: John Benjamins.

Sato, M. (2014). Exploring the construct of interactional oral fluency: Second Language Acquisition and Language Testing approaches. *System, 45*, 79-91.

Sauro, S. (2012). L2 performance in text-chat and spoken discourse. *System, 40*, 335-348.

Segalowitz, N. (2000). Automaticity and attentional skill in fluent performance. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 200-219). Michigan: University of Michigan Press.

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

Skehan, P. (1998). *A cognitive approach to language learning.* Oxford: Oxford University Press.

Skehan, P. (2003) Task-based instruction. *Language Teaching, 36*(1), 1-14

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics*, *30*(4), 510-532.

Skehan, P., & Foster, P. (2012). Complexity, accuracy, and fluency and lexis in task/based performance: A synthesis of the Ealing research. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency* (pp. 199-220). Amsterdam/Philadelphia: John Benjamins.

Spoelman, M., & Verspoor, M. (2010). Dynamic patterns in development of accuracy and complexity: a longitudinal case study in the acquisition of Finnish. *Applied Linguistics, 31*(4), 532–553.

Tavakoli, P. (2016). Fluency in monologic and dialogic task performance: Challenges in defining and measuring L2 fluency. *International Review of Applied Linguistics in Language Teaching, 54(2),* 133-150.

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. 239–277). Amsterdam/ Philadelphia: John Benjamins.

Tavakoli, P., Campbell, C., & McCormack, J. (2015). Development of speech fluency over a short period of time: effects of pedagogic intervention. *TESOL Quarterly*, 1-25.

Thewissen, J. (2013). Capturing L2 accuracy developmental patterns: Insights from an error-tagged EFL learner corpus. *The Modern Language Journal, 97*, 77–101.

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* (pp. 221-245). Amsterdam/Philadelphia: John Benjamins.

Vermeer, A. (2000). Coming to grips with lexical richness in spontaneous speech data. *Language Testing*, *17*(1), 65-83.

Vercellotti, M. L. (2015). The Development of complexity, accuracy, and fluency in second language performance: A longitudinal study. *Applied Linguistics*, 1-23.

Vyatkina, N., Hirschmann, H., & Golcher, F. (2015). Syntactic modification at early stages of L2 German writing development: A longitudinal learner corpus study. *Journal of Second Language Writing, 29,* 28–50.

Wolfe-Quintero, K., Inagaki, S., & Kim, H. Y. (1998). *Second language development in writing: Measures of fluency, accuracy, & complexity* (No. 17). Honolulu: University of Hawaii Press.

ACKNOWLEDGEMENTS

I would like to thank Gabriele Pallotti and the editors who all gave valuable feedback on previous versions of this chapter.

BIOGRAPHICAL NOTE

Marije Michel (PhD, University of Amsterdam) is assistant professor at Lancaster University (UK). Her research focuses on task-based performance, processing and assessment in second language acquisition and in particular the role of task complexity. Recently, she started using eye-tracking methodology to investigate attentional processes during second language (online) writing behaviour.