Applied Linguistics 30/4: 555–578

doi:10.1093/applin/amp044 Advance Access published on 20 November 2009

Towards an Organic Approach to Investigating CAF in Instructed SLA: The Case of Complexity

JOHN M. NORRIS and LOURDES ORTEGA

University of Hawai'i at Mānoa

In this article, we examine current practices in the measurement of syntactic complexity to illustrate the need for more organic and sustainable practices in the measurement of complexity, accuracy, and fluency (CAF) in second language production. Through in-depth review of examples drawn from research on instructed second language acquisition, we identify and discuss challenges to the evidentiary logic that underlies current approaches. We also illuminate critical mismatches between the interpretations that researchers want to make and the complexity measures that they use to make them. Building from the case of complexity, we point to related concerns with impoverished operationalizations of multidimensional CAF constructs and the lack of attention to CAF as a dynamic and interrelated set of constantly changing subsystems. In conclusion, we offer suggestions for addressing these challenges, and we call for much closer articulation between theory and measurement as well as more central roles for multidimensionality and dynamicity in future CAF research.

INTRODUCTION

Fundamental to research in several domains of second language acquisition (SLA) are measures that gauge the three traits of complexity, accuracy, and fluency (CAF) in the language production of learners (including second, foreign, or heritage languages, henceforth L2). These measures typically take the form of ratios, frequencies, or formulas. The global estimates they yield when applied to oral and written L2 data are thought to be reflective of theoretically important constructs which are operationalized via qualities of L2 production, such as subordination for syntactic complexity, variety of word types for lexical complexity, errors for accuracy, or pauses for fluency. Over the past decade, measures of (syntactic and lexical) CAF have been subjected to initial scrutiny. For example, in the context of L2 writing research, Polio (2001) has inventoried measures and examined analytical challenges they present; Wolfe-Quintero *et al.* (1998) have inspected the repeated sampling reliability and concurrent validity of many of these metrics; and Ortega (2003) has provided some benchmarks for interpreting the group

values they yield. A thorough study of the reliability of one lexical measure has been reported by McCarthy and Jarvis (2007), and an interesting discussion of fluency as a construct has been offered by Freed (2000). With the analysis of oral data in mind, Ortega (1999), Skehan (2003), and Robinson and Ellis (2008) have also discussed the issue of whether general estimates or specific measures are more appropriate for SLA research use. These early efforts to synthesize and evaluate what we know about current practices in the use of L2 CAF measures have mostly pointed at the need to improve their use in research, in the traditional sense of enhancing their comparability, reliability, and validity.

In this article, we hope to promote a more considered approach that can help the field to evaluate the evidentiary logic that underlies current uses of CAF measurement. We employ the acronym *CAF* to refer to this conceptual triad, as proposed by the editors of this special issue. We find it memorable in its serendipitous homophony with *calf*, the term used for the offspring of bovine species in the tender age of development between birth and weaning. Playing off of this similarity, we propose an organic approach to CAF which responds to several major challenges and leads to a measurement practice which can inform, rather than confuse, SLA research. First, a major challenge is that, like the similarly named bovine phenomenon, complexity, accuracy, and fluency are each quite complex subsystems with multiple parts, and trying to get a good look at all of the elements that constitute any one of these constructs is a major measurement endeavor. A second challenge is that complexity, accuracy, and fluency qualities (like calves) are developmental in nature, growing, and changing all the time; indeed, CAF as a whole represents a dynamic system of loosely related phenomena that interact in often unpredictable ways. As such, capturing development of multiple subsystems over time, and in relation to each other, is much more difficult than measuring unitary and static aspects of the language, thereby placing heretofore unrealized demands on measurement.

In response to these challenges, we would like to propose here that measurement practices in relation to CAF must become considerably more organic, in the sense that they need to capture the fully integrated ecology of CAF development in specific learning contexts over time, so as to help us understand how and why language develops or not within them. We also suggest that, if an ultimate goal of the CAF agenda is for research to feed back into generalizable understandings about L2 learning, then our measurement practices will have to become much more sustainable, in the sense that what we measure in our individual primary studies will need to be understood in relation to what is being measured in other studies of CAF in other learning contexts. The case of complexity that we explore in this article offers a telling point of departure en route to improved practice, not least because it encapsulates all of the challenges for a measurement practice that is capable of increasing understanding.

An organic model of measurement for understanding

Arguably, the overarching purpose in using CAF measures is to shed empirical light on how the L2 develops, by documenting what parts of the interlanguage system change as acquisition unfolds, in what ways anticipated change proceeds, and perhaps why sometimes not much change seems to take place. Further, for scholars who work in the area of instructed SLA research, 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. In other words, instructed SLA researchers seek to understand phenomena that make a difference in teaching and learning, first and foremost.

What role does measurement play as researchers endeavor to meet these goals? We have proposed elsewhere that, when done well, measurement provides the empirical link between researchable phenomena and the theoretical claims researchers want to make about such phenomena (Norris and Ortega 2003). There are both conceptual and procedural requirements for attaining this ideal, stemming from a process of research that we conceptualize as centered around *interpretation*. This process is depicted in Figure 1, which shows the relationship between interpretations—the target of measurement—and the conceptual and procedural components that influence the extent to which interpretations are warranted.

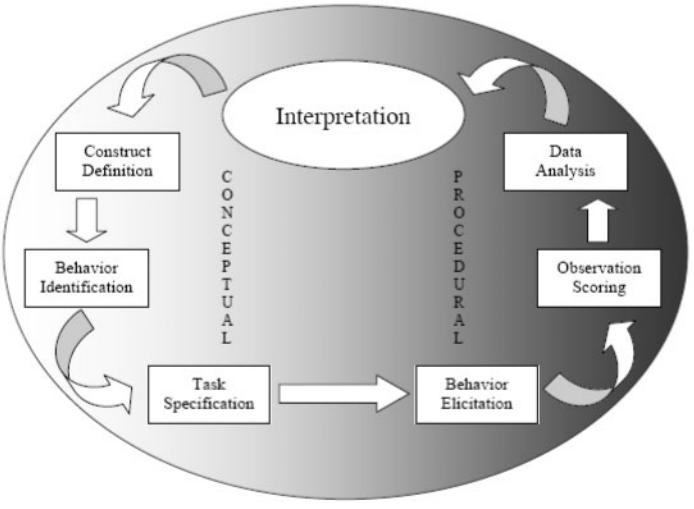

Figure 1: The measurement process
Note: From Norris and Ortega (2003: 720), reprinted with permission

Measurement should help researchers to consistently elicit and observe; it should provide a theory-based framework for interpreting; and it should establish a basis for replicating and accumulating empirical understandings about important constructs. In the case at hand, the interpretation-centered research process should illuminate the constructs of syntactic and lexical CAF in their relationships with and contributions to instructed L2 development. Measurement for understanding should help us to make sense of what we are observing in trustworthy and theoretically meaningful ways.

Couched within this conceptual framework of interpretation-centered measurement, in the remainder of the article, we take one of the constructs in the CAF triad, syntactic complexity, as a test case that will allow us to examine the interface between CAF measurement practice and the theoretically motivated interpretations about instructed SLA that researchers are trying to make.

UNDERSTANDING WHAT WE MEASURE AND WHY

Syntactic complexity is measured across language-related fields by means of a rather wide variety of metrics with a long history, as summarized in Table 1. The syntactic complexity measures most commonly employed across language-related fields are based on *length* and are calculated by dividing words (or morphemes for early childhood languages, and characters for logographic languages) by a chosen production unit. Length-based measures are widely used in child language acquisition (Brown 1973) and in the analysis of L1 writing by school-aged children and adolescents (Hunt 1965; Loban 1976; Scott 1988). The same suite of length measures is popular in SLA. Another class of syntactic complexity metrics includes those that measure amount of *subordination*, and they are computed by counting all clauses and dividing them over a given production unit of choice, yielding, for example, mean number of clauses per T-unit or c-unit. An additional such measure, mean number of clauses per Analysis of Speech unit or AS-unit, is specific to SLA research and has been used frequently since the AS-unit was proposed as an improved option for oral discourse segmentation by Foster *et al.* (2000). Some SLA authors have also employed mean number of dependent or subordinate clauses per total clauses. Amount of *coordination* was proposed by Bardovi-Harlig (1992) as a metric that might be potentially more sensitive than subordination measures, when complexification must be captured at incipient levels of L2 competence. Finally, a rather different definition of syntactic complexity underlies a variety of formulas that have been devised in other fields in order to capture the *variety, sophistication, and acquisitional timing of grammatical forms* used in production, such as Loban's (1976) Elaboration Index or Scarborough's (1990) Index of Productive Syntax. In SLA, variations of this approach include Pienemann's (1998) Rapid Profile assessment system. However, researchers who investigate the relationship between CAF and instructed SLA have not explored these options formally,

Table 1: Main syntactic complexity measures available for research

Central focus of calculation	Measures	Key illustrative references	Comments
Length (in words, morphemes, characters, etc.)	Mean length of utterance Mean length of T-unit Mean length of c-unit Mean length of clause	Brown (1973) Hunt (1965) Loban (1976) Scott (1988)	These measures are easy to calculate and have a long history of use across language-related fields. Most have been employed extensively in SLA as well.
Amount of subordination	Mean number of clauses per T-unit Mean number of clauses per c-unit Mean number of clauses per AS-unit Mean number of dependent or subordinate clauses per total clauses	Elder and Iwashita (2005) Skehan and Foster (2005) Michel *et al.* (2007)	Also easy to calculate, they have increasingly become the measures of choice in SLA research.
Amount of coordination	Coordination Index	Bardovi-Harlig (1992)	Suggested in SLA as an alternative to subordination measures for low-level proficiencies. Not very often employed.
Variety, sophistication, and acquisitional timing of grammatical forms used in production	Elaboration Index Index of Productive Syntax	Loban (1976) Scarborough (1990)	Not used in SLA. They typically involve gauging the frequency of a range of weighted structures. The structures included and their weighting of complexity can be motivated in: (i) empirical findings about emergence sequences attested in developmental data; and (ii) empirically driven difficulty rankings of production or judgements.
Total frequency of use of certain forms considered to be sophisticated	Raw tallies of certain verbal morphology (e.g. tensed forms, passive voice), classes of verbs (e.g. imperatives, auxiliaries, conditionals, modals), syntactic structures (e.g. comparatives, infinitival sentences, conjoined clauses, *wh*-clauses), etc.	Ellis and Yuan (2005); Robinson (2007)	Used in SLA only, probably as an easier to calculate approximation to the sophistication measures employed in other fields (see preceding row).

although some (e.g. Ellis and Yuan 2005; Robinson 2007) have counted the raw frequency of certain grammatical forms thought to be more sophisticated, less frequent, or later acquired, such as modals, passives, or infinitival phrases.

Clearly, then, a variety of measures has been employed to operationalize the construct of syntactic complexity. Yet, as we show in the following sections, researchers have not done sufficient thinking about what we are measuring or why.

Not all syntactic complexity metrics are equal: redundancy and distinctness

It is important to recognize that the various measures outlined in Table 1 are not all created equal. Instead, some of them redundantly measure exactly the same thing, whereas others gauge distinct qualities and dimensions of what can only be understood as a multidimensional construct of syntactic complexity, depending on the sources for complexification that they are able to tap.

Let us first examine the argument that some metrics are redundant and measure exactly the same thing. A good case in point is the family of measures which address subordination in Table 1 and are commonly utilized in L2 research on CAF. What these indices have in common is that they all feature clauses (or subordinate or dependent clauses) in the numerator. These metrics are all equivalent, regardless of the denominator of choice, in that they all tap complexification as a phenomenon of subordination exclusively. The only way in which values for this family of measures would show an increase is when more subordinate or dependent clauses are produced. For example, mathematically speaking, the measure of mean number of dependent or subordinate clauses per total clauses is a faithful replication of mean number of total clauses per multi-clausal unit, only that the ratio is calculated at a level of analysis down the syntactic continuum of dependent/subordinate clause unit → clause → multi-clause unit. Which of these subordination measures that the researchers choose to employ should depend only on which unit of discourse segmentation they consider more appropriate for the data at hand. For example, the utterance or the AS-unit may be more appropriate for dialogic oral data, which contain many nonsyntactic segments. The T-unit, on the other hand, may be ideal for intermediate or advanced written data, which are typically produced in full clauses and sentences. The c-unit, perhaps, can be considered more appropriate for data which are likely to include many nonsyntactic segments, such as the language produced by low-proficiency learners, or with data elicited in oral and/or dialogic contexts involving low levels of formality, and when no measurement of pauses will be performed on oral data (something that segmentation into AS-units requires). In sum, using any one of the measures available in the subordination family listed in Table 1 is sufficient to gauge complexification that is achieved by means

of subordination, and using more than one together for the analysis of the same data would be redundant. When such overlap is the case, it is pointless to employ them in the same study. Redundancy can also cause psychometric trouble, given that most CAF researchers resort to multivariate analyses of variance for their inferential tests, and it is advisable to include only variables that represent independent traits and do not correlate highly with each other. It is for this reason that Tabachnick and Fidell (1996) recommend estimating multicollinearity and singularity of the dependent measures prior to conducting multivariate analyses (see particularly pp. 512–14).

On the flip-side of recognizing that there is redundancy in some metrics, we must also realize that different measures in Table 1 are able to index distinct sources of complexification. Specifically, within the length-based measures listed in Table 1, all the metrics that entail a denominator that is potentially multiple-clausal in scope (i.e. utterance, T-unit, c-unit, and AS-unit) can become longer (e.g. through L2 development) in several possible ways that cannot be determined by inspecting the numerical results that these particular length measures yield. Thus, the addition of subordinate clauses can lengthen an utterance or T-unit, but so can adding adjectives and prepositional phrases that pre- or postmodify nouns, or adding nonfinite verb phrases that modify other elements via nonsubordinating clausal means, or other possibilities. It is for this reason that the mean length of a potentially multi-clausal production unit can only be interpreted as a global or generic metric of linguistic complexity: such measures index overall syntactic complexity. On the other hand, and in stark contrast, clause length is unaffected by variations in the amount of subordination exhibited in production. When the average length of all finite clauses (counted regardless of their status as independent, dependent, or subordinate) is calculated, any increases can only result from the addition of pre- or postmodification within a phrase (via adjectives, adverbs, prepositional phrases, or nonfinite clauses) or as a result of the use of nominalizations, or the process of reduction of clauses into phrases which help to condense information (see the discussion of grammatical metaphor and the synoptic style below). Thus, mean length of clause is radically different from the other length-based measures listed in Table 1. Despite its sharing a superficial similarity with the other length-based calculations, clause length taps a more narrowly defined source of complexification; it must be considered a specific measure that taps complexification subclausally, that is, at the phrasal level. We return to this point later and consider what this observation means for a theory of the role of syntactic complexity in L2 development.

The discussion thus far has led us to identify three measureable subconstructs in syntactic complexity: (i) complexity via subordination, which we argue is measured by any metric with clause (or subordinate or dependent clause) in the numerator; (ii) overall or general complexity, which we argue is measured by any length-based metric with a potentially multiple-clausal unit of production in the denominator; and (iii) subclausal complexity via phrasal elaboration, which we argue is measured by mean length of clause.

Two other distinctly measurable sub-constructs of syntactic complexity must be added, although neither has been utilized much in SLA thus far. One is the measurement of increased clausal complexification achieved not via subordination but via coordination, quite relevant for data at initial levels of L2 development as proposed by Bardovi-Harlig (1992). Perhaps, this metric has been an underutilized tool simply because the tendency in most SLA research is to investigate intermediate levels of proficiency. The other measurable subconstruct, which has been employed primarily in other fields, is complexity defined as the variety, sophistication, and acquisitional timing of forms produced (e.g. Scarborough 1990). In lieu of such a metric for L2 research, some (but not most) CAF researchers have tallied the raw frequencies of certain forms that are thought to be more sophisticated or difficult. In the ultimate instance, however, it will be necessary to aim for the creation of L2 indices that offer well-theorized and potentially better specified operationalizations of variety, sophistication, and acquisitional timing of particular forms, based on more fine-tuned L2 developmental findings.

In sum, we have argued that in the measurement of syntactic complexity, there are some metrics that are redundant if used together, because they tap the same measurable dimension of the construct and, conversely, there are other measures that are distinct and complementary and thus can be best used and interpreted together, because they tap different dimensions of complexity. It should be noted that the argument for construct multidimensionality can be and has been made for the other CAF constructs as well. Indeed, for fluency, Skehan (2003, this issue) and Tavakoli and Skehan (2005) have pointed out that there may be three distinct sub-constructs of fluency: (i) breakdown fluency, which is measured via silence-related metrics; (ii) speed fluency, which is tapped by rate- and time-related measures; and (iii) repair fluency, which is gauged by means of self-correction measures. Nevertheless, there seems to be little awareness among instructed SLA researchers who use CAF that the kit of CAF measures that they have at their disposal is not composed of mutually interchangeable or equivalent metrics within each of the three elements in the triad.

Arguments for measuring complexity multidimensionally

There are both theoretical and empirical justifications for our claim that syntactic complexity must be measured multidimensionally. A major theoretical justification can be found in systemic functional linguistics (Halliday and Mathiessen 1999). This theory of language posits that development proceeds from: (i) the expression of ideas first by means of mostly parataxis (i.e. coordination) or the sequencing of self-standing words, sentences, and clauses; through (ii) an expansion by which hypotaxis (i.e. subordination) is added as a resource to express the logical connection of ideas via grammatically intricate texts; to finally (iii) the emergence of and reliance on grammatical metaphor (achieved through nominalization, among other processes), which

leads to advanced language that actually exhibits lower levels of subordination or grammatical intricacy but much higher levels of lexical density and more complex phrases (as opposed to more clauses).

An example discussed by Halliday and Martin (1993/1996: 31–41) helps to illustrate this theoretical proposal. They compared the seven-word complex T-unit shown in (1) with the six-word equivalent phrase shown in (2):

(1) *Darwin thought that species gradually became more complex.*
(2) *Darwin's gradual rise to mounting complexity . . .*

In (1), the authors paraphrased (2), which is a phrase attested in the writing of the late biologist Stephen Jay Gould. The same content is expressed by means of subordination in (1), but by means of grammatical metaphor in (2), specifically via nominalization (*became*-verb → *rise*-noun; and *complex*-adjective → *complexity*-noun) and other grammatical class substitutions, such as making an adverb into an adjective (*gradually* → *mounting*). In addition, the version in (2) can be combined with many other phrases in order to express even denser and longer ideas (e.g. *Darwin's gradual rise to mounting complexity is a remarkable intellectual achievement*). The difference between the two options is also described in systemic functional linguistics as a change from a dynamic style to a synoptic style of expression, which occurs whenever 'what might be construed as a combination of interdependent clauses . . . is reconstrued as edifice of words and phrases . . . ' and 'The meaning comes to function . . . at a lower rank in the grammar – at the ranks of group/phrase and word, instead of at the rank of clause' (Halliday and Martin 1993/1996: 39). Individuals learning a (first or second) language are expected to change along these lines as they become more sophisticated in their linguistic competencies and as they grow more capable of handling written and academic-formal registers. This prediction is posited to be universal by systemic-functional linguists, although more cross-linguistic validation seems necessary (see discussion in Ortega 2003: 514–15).

If the theoretical model of language development as an expansion from dynamic to synoptic styles is correct, then coordination can be expected to be the most indicative source of complexification at beginning levels of development (as claimed by Bardovi-Harlig 1992), and subordination should be a useful and powerful index of complexification at intermediate levels. However, subordination's role should subside at even higher levels of development in favor of greater use of phrasal-level complexification, which should become the pervasive means by which syntactic complexity is achieved at the most advanced levels of language development and maturity. Working from a corpus linguistic perspective that is theoretically agnostic towards Hallidayan grammar, Biber (2006) has provided extensive empirical evidence for precisely such patterns of use in mature L1 English academic registers.

Returning to our discussion of which family of measures taps which dimensions of complexity, it turns out that the Coordination Index ought to have a much great predictive power when measuring syntactic complexity at

beginning levels of L2 development, while any of the subordination measures ought to be of greater value when measuring the construct at intermediate and upper-intermediate levels. Finally, mean length of clause (the only measure to date that taps complexification at the subclausal or phrasal level) ought to be most predictive at an advanced point in development, when processes of grammatical metaphor begin to unfold and more synoptic styles emerge in the repertoires of high-proficiency L2 learners and users. It follows, then, that it will be wise to measure all three dimensions of complexity in the same data, and this will require minimally the combined use of one measure from each of the three families in the same study.

The justification for our claim that complexity (and probably all other) CAF constructs must be measured multidimensionally is not only theoretical, but also empirical. For example, after reviewing the findings available for school-aged L1 writing at the time, Scott (1988: 58) concluded that subordination (expressed as the average number of clauses per T-unit in this case) leveled off by eighth grade, even though mean length of T-unit kept increasing beyond that point. In two investigations of L2 written data involving comparisons of only two levels of college ESL learners, less subordination for the more advanced group was reported by Perkins (1980) and Bardovi-Harlig and Bofman (1989). Another three L2 studies also involving college students but each including multiple-level comparisons (for German: Cooper 1976; for French: Monroe 1975; for English: Flahive and Snow 1980) reported a leveling off or decrease in subordination towards the upper proficiency levels, which was located at the third year of college juncture for the German and French studies and at the fifth of six levels in the L2 English study. While all these investigations were cross-sectional, new empirical evidence supporting exactly the same interpretations is also emerging from longitudinal German L2 data (Byrnes *et al.* in press).

The pattern of results that we have discussed here has been noted previously by Wolfe-Quintero *et al.* (1998: 73) and Ortega (2003: 514–15). However, empirical studies of CAF have not pursued this line of reasoning. How exactly, then, do SLA researchers select among these diverse subconstructs, and among the available individual metrics, when they wish to measure complexity in their studies? We turn to this question in the next section.

CURRENT MEASUREMENT PRACTICES IN CAF-BASED RESEARCH

Extending our use of syntactic complexity as an illustrative case of current CAF practices, we perused the choices made regarding syntactic complexity metrics across 16 empirical studies that have measured CAF in the context of research on task-based language learning. The results are summarized in Table 2, using the proposed distinct families of the multidimensional construct

Table 2: Complexity dimensions and measures used across 16 studies in recent task-based language learning research

	Overall complexity	Complexity by subordination	Phrasal complexity	Variety of forms
Albert and Kormos (2004)		Clauses per AS-unit		
Elder and Iwashita (2005)		Clauses per c-unit		
Ellis and Yuan (2005)		Clauses per T-unit		Frequency of: Tensed forms, Modals, Passive forms
Gilabert (2007)		S-nodes per T-unit		
Ishikawa (2007)	Mean length of T-unit[a]	S-nodes per T-unit Clauses per T-unit Dependent clauses per clause S-nodes per clause	Mean length of clause[a]	
Iwashita *et al.* (2001)		Clauses per c-unit		
Kawauchi (2005)	Mean length of T-unit	Clauses per T-unit Frequency of subordination		
Kuiken and Vedder (2007)		Clauses per T-unit Dependent clauses per clause		
Lambert and Engler (2007)		S-nodes per AS-unit		
Michel *et al.* (2007)		Clauses per AS-unit Subordinate clauses per total clauses		
Robinson (2007)	Mean length of c-unit[a] Mean length of turn	Clauses per c-unit		Frequency of: Infinitival phrases, Conjoined clauses, *Wh*-clauses
Sangarun (2005)		Clauses per T-unit S-nodes per T-unit		Frequency of: Imperatives, Auxiliaries, Comparatives, Conditionals
Skehan and Foster (2005)		Clauses per AS-unit		
Storch and Wigglesworth (2007)	Mean length of T-unit[a]	Clauses per T-unit Subordinate per dependent clauses Subordinate clauses per T-unit Subordinate per total clauses	Mean length of clause[a]	
Tavakoli and Foster (2008)	Mean length of AS-unit	Clauses per AS-unit		
Tavakoli and Skehan (2005)		Clauses per AS-unit		

Note: The studies included in this table are all empirical, quantitative studies of task-based language learning that employed CAF measures. They were identified by perusing three collections (Ellis 2005; García Mayo 2007; Robinson and Gilabert 2007) and all empirical studies published in *Language Learning* between 2000 and 2008, both years inclusive.

[a]Employed as a measure of fluency, not complexity.

that we have outlined in the previous section. It can be seen there, first of all, that only six of the 16 studies employed distinct measures for at least two construct dimensions. More specifically, subordination measures were employed in all 16 studies without fail, and by comparison little interest was demonstrated in gauging possible changes in complexification at the phrasal or at the global level. In fact, only two studies used mean length of clause, and even then both claimed to use it as a measure of fluency, not complexity. Likewise, only five of the 16 studies featured any measure of overall or general complexity (i.e. any length-based measure with a potentially multi-clausal denominator), and three used such metrics to purportedly measure fluency, not complexity. A second general observation, and a more positive one, is that three studies included among their complexity analyses frequency counts of selected forms, in an attempt to tap complexity as structural variety, sophistication, and acquisitional timing of grammatical forms.

Let us comment first on the fact that only six of the 16 studies employed distinct metrics for at least two measurable construct dimensions. The reduction of complexity to only one of the dimensions is an extended practice in L2 research, not only in task-based investigations. In the L2 writing domain, however, mean length of T-unit seems to be the single most employed complexity measure (Wolfe-Quintero *et al*. 1998; Ortega 2003), and thus global complexity seems to be prioritized over other dimensions of the construct. In contrast, in research on task-based language learning, as Table 2 suggests, the single dimension that seems to be prioritized is complexification by subordination. The exclusive reliance on subordination measures is worrisome, for the theoretical and empirical reasons we previously discussed. At best, positing that elaboration occurs by subordination across all levels of development is too limited a view of complexity. At worst, on the basis of subordination measures alone we may completely misinterpret whether an increase or a decrease is indicative of a positive or negative change in performance, because a decrease in subordination at the highest levels of proficiency may be related to an increase in the overall complexity of the language performance.

It is also remarkable that in six cases researchers decided to employ what can be considered equivalent or redundant metrics that tap the same kind of complexity: once again, subordination. Not surprisingly, most of them reported that each of the subordination measures yielded the same results, all supporting the hypotheses concerning complexity (Kawauchi 2005; Ishikawa 2007) or all not showing the expected effects (Kuiken and Vedder 2007; Michel *et al*. 2007; Storch and Wigglesworth 2007). The only exception is Sangarun (2005), who found a predicted statistically significant difference in complexity with the S-nodes per T-unit measure but not with clauses per T-unit. It might be that calculations with the S-node in the numerator, which include both finite and nonfinite clauses, show a greater sensitivity for measuring small differences in complexity at relatively low levels of proficiency, such as high school students of English as a foreign language in Thailand. In any case, while some may want to interpret the overall consistency in most of these studies as a sign

of the robustness of the given findings, it would be more accurate to say that such evidence is redundant rather than robust.

The second salient observation from Table 2 is the inclusion in some studies of frequency counts of selected forms among the complexity analyses. This pattern can be taken as a sign that CAF researchers are increasingly interested in measuring complexity as structural variety and sophistication. Robinson (2007) has mentioned the possibility of moving forward in this direction if in the future L2 researchers were willing to adopt or adapt Scarborough's (1990) Index of Productive Syntax for L2 data, and Granfeldt and Nugues (2007) have developed such a weighted and broad-structure approach for L2 French in their Direkt Profil. However, attention also needs to be directed to devising measures that include a wide range of developmentally ranked structures regardless of their status as target-like or nontarget-like, so as to help researchers characterize L2 production that ranges along the full developmental continuum. Recent explorations in this vein have included efforts at devising polytomous or weighted scoring systems that allow for the coding of L2 production in a given grammar area of interest in a developmentally motivated fashion. Thus, based on well-known findings about the L2 acquisition of tense and aspect, Révész (2009) scored learners' use of the past progressive on oral tasks as 0 (if no progressive was attempted), 1 (if a bare gerund was produced), 2 (if present progressive was produced), or 3 (if the target-like form of the past progressive was used). A similar system was devised by Mochizuki and Ortega (2008) for the developmental measurement of relative clause production, drawing also on well-known L2 findings in this area. Efforts to advance the field's ability to engage in more form-specific and development-sensitive measurement of L2 production deserve attention, as Robinson and Ellis (2008) and Robinson *et al.* (this issue) note. However, much more validation work remains to be done in this area.

HOW DO WE KNOW WHAT WE ARE MEASURING?

As shown in Table 2, many researchers who employ length-based measures to study CAF claim to understand them as measures of fluency, not complexity. This tendency is true of both mean length of a multi-clausal unit of production and mean length of clause. This state of affairs is a new development since the publication of Wolfe-Quintero *et al.*'s (1998) book, which has become a standard citation in this research domain. Wolfe-Quintero *et al.* acknowledged that length-based measures have always been used to measure complexity but declared that they did not agree with this traditional interpretation, proposing that all metrics involving mean length of any production unit should be considered measures of fluency. However, this proposal presented a disconcerting departure from the originally intended use for measures as central to other fields as mean length of utterance (Brown 1973) and mean length of T-unit (Hunt 1967). Contested construct interpretations must be evaluated empirically, rather than being embraced by fiat. Therefore, we first

review the arguments provided by Wolfe-Quintero *et al*. for their proposal and we then examine the results from an unpublished study by Oh (2006) which offer relevant evidence on the issue of whether length-based measures tap fluency or complexity.

The evidence that Wolfe-Quintero *et al*. provided for their construct reinterpretation of length-based measures is scant. First, they argued that such measures must be considered to index fluency, because as an index of complexity they would not be informative enough to tell us how the complexification (i.e. the lengthening) is achieved. This argument is a good reason for using length-based measures of overall or general complexity (which, we argue, are those with a potentially multi-clausal unit of production in the denominator) in conjunction with more fine-grained and specific measures, such as measures of complexity via subordination and via phrasal elaboration. However, methods that afford global, unspecified, or general interpretations may still be legitimate ways of measuring a given construct, just at a different grain size. Moreover, we still may need to have a good measure of overall or general complexity because such a broader measure might be able to capture large-scale or long-term variation that would be missed by finer-grained, more specific metrics.

The only other rationale offered by Wolfe-Quintero *et al*. for their proposal rests with their interpretation of the findings reported by Ortega in her unpublished master's thesis, where a Varimax rotation factor analysis of 10 CAF measures resulted in a clear three-factor structure, representing fluency-plus-complexity, accuracy, and lexical diversity (Ortega 1995, discussed in Wolfe-Quintero *et al*. 1998: 14). The first factor was of interest to Wolfe-Quintero *et al*., because two calculations of speech rate (considered fluency measures) as well as words per utterance and propositions per utterance (both considered length-based complexity measures) loaded very highly and exclusively on Factor 1.

The main results in question were reported in Ortega (1999), but the factor analysis was not included in that publication because it was purely exploratory and conducted with the purpose of identifying potentially redundant variables in order to arrive at a more parsimonious CAF selection that would improve the statistical power of the design. By this process, five of the 10 measures were kept for the main analyses reported in Ortega (1999). In the original study, moreover, Ortega (1995) offered a very different interpretation from that of Wolfe-Quintero *et al*. Far from concluding that length measures must be considered to tap fluency, Ortega cast doubts on the validity of using the utterance (intonationally and pausally defined, as proposed by Sato 1990) as the denominator in complexity measures, as was the case in that study. Specifically, she argued that by definition the utterance conflates complexity with fluency and that this confounding problem 'may be aggravated when measuring syntactic complexity in low or intermediate levels of oral proficiency' (Ortega 1995: 118). In other words, learners may be able to produce more or less sophisticated or complex language, a question of complexity,

and at the same time (particularly low-proficiency learners) may or may not be able to utter them in one stream of speech, a question of fluency. In fact, we believe that having the utterance as the denominator in a length-based measure is probably identical to the metric of mean length of run that Tavakoli and Skehan (2005) have identified as a promising option for measuring fluency. They define mean length of run, following Freed (2000), as a 'continuous stream of running speech (measured in words) not interrupted by disfluent pauses or hesitations', and they note this measure is reflective of 'how lengthy the language produced between two pause boundaries is' (p. 255).

Some empirical evidence that the traditional complexity interpretation of length-based measures is valid comes from Oh's (2006) investigation of a corpus of 78 argumentative essays written by international undergraduates in the USA as part of mandatory placement procedures. It is important to note that the students' general proficiency was high, as demonstrated by Internet-based TOEFL scores that ranged from 61 to 100 (i.e. 500 to 600 TOEFL paper-based scores). Oh was interested in determining how well fluency would predict the 78 students' placement into one of two writing levels available in the program: (i) a credit-bearing, upper-level undergraduate ESL writing course ($N = 43$), equivalent to the freshman composition course that US students at the same university must complete; or (ii) an intermediate-level course ($N = 35$) that must be completed before taking the upper-level one. Following Wolfe-Quintero *et al*. (1998), she calculated six 'fluency' measures, with a simple intercoder agreement ratio that was higher than 0.95 for all segmentation units involved. Three of the measures tapped raw frequencies reflecting the total amount of language produced, or what can be thought of as productivity (total number of words, T-units, and clauses). Another measure was a ratio that reflected the rate of written production (words per total minutes spent writing). The remaining two measures (mean length of T-unit and clause) were length-based measures that we consider to measure syntactic complexity (cf. Table 1), but that have been argued by Wolfe-Quintero *et al*. to tap fluency. Oh submitted the results to a rotated Varimax factor analysis with a minimum Eigenvalue set at 1.00, which yielded a sufficiently reliable Kaiser–Meyer–Olkin and Bartlett's test result of 0.55 at $p = 0.000$.

The basic factor loadings are listed in Table 3. It is clear that in these data four of the six measures loaded very highly together on the first factor and, in contrast, mean lengths of T-unit and clause loaded very highly together on the second factor. Each factor also explained a considerable proportion of unique variance observed. Note that the first factor included measures that were computed by raw frequency counts as well as by a ratio calculation. Thus, the possibility of a superficial commonality, like the length-based calculations, explaining the results is clearly inadequate. Rather, it can be argued that the first factor reflects fluency, whereas the second one reflects complexity. If this interpretation is correct, then the analyses support the traditional construct definition of mean lengths of T-unit and clause as complexity measures.

Table 3: Rotated factor matrix in Oh (2006)

	Factor	
	1	2
Number of words	**0.96**	0.17
Number of clauses	**0.90**	−0.33
Number of T-units	**0.84**	−0.46
Words per minute	**0.83**	0.13
Number of words per T-unit	−0.02	**0.92**
Number of words per clause	−0.04	**0.89**
Eigenvalue	3.30	1.83
Percentage of variance	55.08	30.43
Cumulative percentage	55.08	85.51

Note: Reproduced with author's permission from Oh (2006), Table 17.

In the end, as new findings like Oh's (2006) emerge to shed empirical light on proposals such as that made by Wolfe-Quintero *et al*. (1998), they remind us that we really need to establish interpretation-centered warrants for what our measures purportedly are measuring. The nature of the units that are used, the content contribution of numerator and denominator, and the quality of the information the values may provide, must all be carefully considered, and these considerations must be supported by theory and empirical evidence. Cumulative empirical investigation of the distinctiveness and interrelationships among CAF constructs will be of the essence if we are to understand minimally whether we are observing complexity, fluency, or some other phenomenon when we measure L2 production data.

UNDERSTANDING CAF INTERRELATIONSHIPS: DYNAMIC PREDICTIONS FOR DYNAMIC CAF

In this last section, our focus expands beyond complexity and looks to its relationship with other complexity, accuracy, and fluency, constructs. This move will enable illustration of the interdependence and dynamicity of the three traits of complexity, accuracy, and fluency. Our basic claim is that we cannot expect constant predictive value among complexity, accuracy, and fluency, and that instead we must model multivariate interactions in our measurement-based observations.

We turn first to results reported by Norris (1996) to illustrate dynamicity in the interrelationship among CAF constructs. The study involved an extensive CAF analysis of an L2 German corpus of 44 taped oral proficiency assessments. The tapes had been recorded in 1995, when the German Speaking

Test (Center for Applied Linguistics 1995), a Simulated Oral Proficiency Interview, was pilot tested with 44 volunteer examinees recruited in several universities and high schools in the USA. The performances were rated in-house at CAL according to the oral proficiency guidelines of the American Council on the Teaching of Foreign Languages (ACTFL 1985), and the data were subsequently transcribed and analyzed by the researcher. The proficiency ratings of the samples spanned eight levels, from Novice-Mid to Superior on the ACTFL Guidelines scale, a very wide range of overall L2 proficiency. The range can be put into a curricular perspective if one remembers that students who enroll in German first- and second-year courses in college typically are rated from Novice-Low to Intermediate-Mid, whereas ratings of Intermediate-High to Advanced are more typical of students in their third and fourth years of college German, and ratings of Superior are quite uncommon (Norris and Pfeiffer 2003).

Norris employed a large battery of measures tapping syntactic and lexical CAF, including the following indices among others: mean lengths of T-units and clauses; type-token ratio, lexical sophistication/uniqueness, non-German lexis; accuracy rates with rules of German word order; words per second, number and length of pauses, and total words. Employing exploratory factor analysis and discriminant analysis among other techniques, Norris sought to identify relationships among the measured CAF variables and the extent to which any or all of them were associated with proficiency differences. Turning to the factor analyses, conducted on data from long and short forms of the test, multiple discrete factors were identified for measures of: (i) syntactic complexity, (ii) lexical complexity, (iii) grammatical accuracy, and (iv) fluency. That is, these measures behaved independently enough to be identifiable as separate factors in relation to each other (so, identifying C, A, and F made some sense in these data). Given their distinctiveness in the oral performance data, Norris was then able to examine how these features of CAF distinguished among oral performances rated at differing proficiency levels, using discriminant and descriptive analyses.

A first finding was that a single accuracy measure, percentage correct use of the verb-end word order rule (Meisel *et al*. 1981), was able to distinguish with 100 per cent accuracy between ratings of Intermediate-Mid and below versus Intermediate-High and above. Second, it became apparent that distinct features of CAF played a greater or lesser role in distinguishing performances in different proficiency ranges. Thus, at the lower end of the proficiency scale, fluency variables (speech rate, total speech, and number/length of pauses) proved effective at distinguishing between each of the Novice-Mid through Intermediate-High levels, while these variables were not the most effective at distinguishing among higher proficiency samples. In contrast, lexical complexity variables (as measured by lexical uniqueness and use of non-German lexis) proved to be the best indicators of differences between Intermediate-High to Superior rated samples, while such measures were poor at distinguishing among the lower proficiency levels.

Similarly, in the study previously discussed, Oh (2006) also performed discriminant analyses in order to explore the predictive value of clusters of the six measures. She entered the first identified factor (four measures interpreted as 'fluency'), the second factor (two measures interpreted as 'complexity'), and all six measures together. The results are shown in Table 4.

We interpret the outputs shown in Table 4 as follows. For these data and in this university context, the four fluency measures alone were equally useful in predicting placement into either of the two levels. That is, fluency offered a good overall predictive value of 67.9 per cent of all 78 cases and only a modest advantage of 4.1 per cent extra predictive value in favor of the upper course placement. When the two complexity measures alone were used, on the other hand, the predictability of placing writers into the lower level course was clearly superior, with an advantage of 12.8 per cent better prediction in favor of lower course placement, and this was true for 68.9 per cent of the 35 cases, but the predictability of placing writers into the upper-level course was not as strong as with the fluency measures. Finally, the combined use of all four fluency and two complexity measures was most useful for the accurate prediction of placement into the upper level, offering an advantage of 18.5 per cent better prediction for the upper level writers,

Table 4: Prediction of group membership according to three discriminant analyses in Oh (2006)

	Placement level	Lower course	Upper course	Total
Predicted group membership with Factor 1 ('Fluency')[a]				
Raw Sample Count	Lower course	23	12	35
	Upper course	13	30	43
Percentage	Lower course	**65.7**	34.3	100.0
	Upper course	30.2	**69.8**	100.0
Predicted group membership with Factor 2 ('Complexity')[b]				
Raw Sample Count	Lower course	24	11	35
	Upper course	19	24	43
Percentage	Lower course	**68.6**	31.4	100.0
	Upper course	44.2	**55.8**	100.0
Predicted group membership with all six measures combined[c]				
Raw Sample Count	Lower course	22	13	35
	Upper course	8	35	43
Percentage	Lower course	**62.9**	37.1	100.0
	Upper course	18.6	**81.4**	100.0

Note: Table adapted with author's permission from Oh (2006), Tables 26, 27, and 28
[a]67.9 per cent of all original grouped cases correctly classified.
[b]61.5 per cent of all original grouped cases correctly classified.
[c]73.1 per cent of all original grouped cases correctly classified.

and this was true for 81.4 per cent of the 43 cases. Simply put, fluency alone predicted equally well the placement of lower and upper level writers, complexity alone afforded a better predictive power for the lower level writers, and fluency and complexity combined resulted in the best predictions for the upper level writers. This pattern of results suggests that writers exhibiting weaker complexity in their essays tended to be placed in the lower level course, and that it was L2 writers with healthier profiles exhibiting both relative fluency and higher complexity who tended to be placed into the more advanced writing course.

In sum, what these two studies highlight is that different qualities of production, alone or combined, can have lesser or greater predictive value depending on the relative proficiency levels of L2 learners under investigation. That is, qualities of L2 production cannot and should not be considered constant along the CAF developmental continuum, and different CAF traits must serve different interpretive purposes for different proficiency levels. This conclusion becomes apparent when CAF evidence is derived from factor analysis and discriminant analysis, and when developmental and proficiency considerations are included organically in the interpretation of empirical CAF findings. It also holds relevance not only with broad proficiency ranges (as in the Norris study), but even with a sample whose proficiency is rather narrow and towards the higher ends of the continuum (as in the Oh study). Not all traits of CAF will have an equally predictive value for all proficiency levels. Development is a long-term and multifaceted process, and data must be interpreted with an awareness of where along that process the evidence is being collected and analyzed. CAF, it seems, consists of a variety of dynamically related indices which do not all advance hand-in-hand towards an ideally complex, accurate, and fluent performance. Indeed, depending on the proficiency or developmental levels of learners, if we focus only on anticipated changes in one area, we may be missing the really important changes (or lack thereof) going on in another.

CONCLUSION

Many other measurement challenges and observations deserve serious attention, omitted here only for the sake of space. Among them, we point to three in particular: (i) the dire need to assess and enhance the quality of corpora on which evidence about L2 acquisition must be based (Tomasello and Stahl 2004); (ii) the imperative to explore the theoretical status of variability in the study of CAF (Larsen-Freeman 2006; Verspoor *et al.* 2008); and (iii) the need to undertake more longitudinal investigations in the future if we are to obtain a full view of developmental trajectories (Ortega and Byrnes 2008). We would like to conclude with a few suggestions that follow from the arguments we have developed in this article, and that we think will help in using interpretation-centered measurement that enables a better understanding of CAF.

If we really want to understand how and why L2 CAF develops, what would we hope to see in the immediate future? First, it would be advisable to employ measurement practices that engage with the construct reality of multidimensionality. In the case of syntactic complexity, specifically, at a minimum SLA researchers should measure global or general complexity, complexity by subordination, and complexity via phrasal elaboration, as well as possibly coordination if early proficiency data are also included. That will demand the use within single studies of metrics chosen to tap at least overall complexity (e.g. mean length of T-unit), complexity by subordination (e.g. mean number of clauses per T-unit), and complexity by subclausal or phrasal elaboration (e.g. mean length of clause). In addition to seeing a better motivated use of these more global measures, we would also hope that more specific measures will be devised and used for L2 data specifically, as Robinson *et al*. (this issue) argue, but also more developmentally sensitive and interlanguage-based measures that tap complexity defined as structural variety, sophistication, and acquisitional timing. Further, when several complexity measures are used in the same L2 study, an easy and complementary analysis would be to inspect multicollinearity, report how strongly the various dependent variable measures correlate with each other, and strive to achieve singularity (Tabachnick and Fidell 1996). Finally, in studies where a variety of CAF measures are employed, and provided that the sample sizes in question are large enough, it would be useful to employ factor analysis to see which measures contribute to which (and how many) factors, in order to align empirical evidence and theoretical predictions more tightly and to improve the final interpretations made in the study.

At a more general level, what are the requirements for interpretation-centered CAF measurement? We suggest that researchers must engage in a much more organic practice in order to achieve a thorough understanding of CAF as conditioned by the realities of learning contexts. On the one hand, it means that our measurements must provide multivariate, longitudinal, and descriptive accounts of constructs in L2 performance in order to capture the complex, dynamic, and developmental nature of CAF phenomena. On the other hand, it means that measurement will also need to provide learner-, task-, and L2 form-sensitive accounts of the local SLA ecology, given the ways in which these factors moderate the observations we might be making about CAF. The examples presented in this article and in this special issue demonstrate the critical necessity of paying attention to the demands of organic CAF measurement practices. At the same time, we would suggest that researchers will need to adopt much more sustainable practices if our individual research efforts are to contribute maximally to replicable, cumulative, and perhaps even amendable understandings of CAF across study contexts. Along these lines, it will be essential that we (researchers, editors, and publishers) adjust our reporting practices such that we begin to include accurate and complete depictions of measurement tools, data, and analyses (Norris and Ortega 2000, 2006). Furthermore, we would hope to

encourage much greater sharing of L2 learner CAF data and corpora with the public (e.g. CHILDES database, MacWhinney 2000; the Myles and Mitchell's French Learner Language Oral Corpora database, http://www.flloc.soton.ac.uk/index.php), such that claims may be re-examined, phenomena measured in alternative ways, and thereby interpretations about CAF rendered much richer.

Having devoted close attention to complexity and CAF in this article, we would like to end with a cautionary note. In the ultimate analysis, it is illusory to think that what we are measuring in CAF is some kind of universal construct that can be applied across all possible learners and contexts. CAF is always measured for particular purposes in particular settings and with particular developmental targets in mind. These factors must condition why and how we interpret language learning. Just as calves may turn into *toros bravos*, revered deities, or milk cows, so too must we consider carefully just what CAF is intended to become in any learning setting and for the educational, linguistic, and societal cultures that define it. Likewise, in an organic and sustainable approach to instructed SLA research, raising CAF probably is not going to be sufficient. There is certainly much more to language learning and language use than what measures of CAF might account for. We probably need to be careful in our zeal for focusing on CAF that we do not ignore other phenomena essential to a more complete understanding of second language learning.

ACKNOWLEDGEMENTS

An earlier version of this article was presented at the American Association for Applied Linguistics conference in Washington DC, 2008. We thank Leila Ranta, Ute Römer, and Parvaneh Tavakoli for their comments from the audience. We are also grateful to Alex Housen, Folkert Kuiken, four reviewers, and Doug Biber for their helpful suggestions.

REFERENCES

ACTFL (American Council for the Teaching of Foreign Languages). 1985. *ACTFL Proficiency Guidelines*. Author. Also retrievable from: http://pnglanguages.org/lingualinks/LANGUAGELEARNING/OtherResources/ACTFLProficiencyGuidelines/contents.htm

Albert, A. and **J. Kormos.** 2004. 'Creativity and narrative task performance: An exploratory study,' *Language Learning* 54: 277–310.

Bardovi-Harlig, K. 1992. 'A second look at T-unit analysis: Reconsidering the sentence,' *TESOL Quarterly* 26: 390–5.

Bardovi-Harlig, K. and **T. Bofman.** 1989. 'Attainment of syntactic and morphological accuracy by advanced language learners,' *Studies in Second Language Acquisition* 11: 17–34.

Biber, D. 2006. *University Language: A Corpus-Based Study of Spoken and Written Registers*. John Benjamins.

Brown, R. 1973. *A First Language: The Early Stages*. Harvard University Press.

Byrnes, H., H. Maxim, and **J. M. Norris.** 'Realizing Advanced L2 Writing Development in a Collegiate Curriculum: From Outcomes Expectations to Assessment. Monograph Series,' *Modern Language Journal* 94/5, in press.

CAL (Center for Applied Linguistics). 1996. *German Speaking Test Author*.

Cooper, T. C. 1976. 'Measuring written syntactic patterns of second language learners of German,' *Journal of Educational Research* 69: 176–83.

Elder, C. and **N. Iwashita.** 2005. 'Planning for test performance: Does it make a difference?' in Ellis (ed.).

Ellis, R. (ed.) 2005. *Planning and Task Performance in a Second Language*. John Benjamins.

Ellis, R. and **F. Yuan.** 2005. 'The effects of careful within-task planning on oral and written task performance' in R. Ellis (ed.): *Planning and Task Performance in a Second Language*. John Benjamins.

Flahive, D. E. and **B. Snow.** 1980. 'Measures of syntactic complexity in evaluating ESL compositions' in J. W. Oller and K. Perkins (eds): *Research in Language Testing*. Newbury House.

Foster, P., A. Tonkyn, and **G. Wigglesworth.** 2000. 'Measuring spoken language: A unit for all reasons,' *Applied Linguistics* 21: 354–75.

Freed, B. 2000. 'Is fluency, like beauty, in the eyes (and ears) of the beholder?' in H Riggenbach (ed.): *Perspectives on Fluency*. University of Michigan Press.

García Mayo, M. P. (ed.). 2007. *Investigating Tasks in Formal Language Learning*. Multilingual Matters.

Gilabert, R. 2007. 'The simultaneous manipulation of task complexity along planning time and (+/− Here-and-Now): Effects on L2 oral production' in M. P. García Mayo (ed.): *Investigating Tasks in Formal Language Learning*. Multilingual Matters.

Granfeldt, J. and **P. Nugues.** 2007. 'Evaluating stages of development in second language French: A machine-learning approach' in J. Nivre, H.-J. Kaalep, K. Muischnek, and M. Koit (eds): *NODALIDA 2007 Conference Proceedings*. Tartu, Estonia, May 25–26. Available at http://person.sol.lu.se/JonasGranfeldt/publikationer.html. Accessed 15 February 2009.

Halliday, M. A. K. and **J. R. Martin.** 1993. *Writing Science: Literacy and Discursive Power* (Vol. first printed in 1993, reprinted in 1996). Falmer Press.

Halliday, M. A. K. and **C. Matthiessen.** 1999. *Construing Experience through Meaning: A Language-Based Approach to Cognition*. Cassell.

Hunt, K. W. 1965. *Grammatical Structures Written at Three Grade Levels*. National Council of Teachers of English.

Ishikawa, T. 2007. 'The effect of manipulating task complexity along the [+/− Here-and-Now] dimension on L2 written narrative discourse' in M. P. Mayo García (ed.): *Investigating Tasks in Formal Language Learning*. Multilingual Matters.

Iwashita, N., T. McNamara, and **C. Elder.** 2001. 'Can we predict task difficulty in an oral proficiency test? Exploring the potential of an information-processing approach to task design,' *Language Learning* 51: 401–36.

Kawauchi, C. 2005. 'The effects of strategic planning on the oral narratives of learners with low and high intermediate L2 proficiency' in R. Ellis (ed.): *Planning and Task Performance in a Second Language*. John Benjamins.

Kuiken, F. and **I. Vedder.** 2007. 'Cognitive task complexity and linguistic performance in French L2 writing' in M. P. García Mayo (ed.): *Investigating Tasks in Formal Language Learning*. Multilingual Matters.

Lambert, C. P. and **S. Engler.** 2007. 'Information distribution and goal orientation in second language task design' in M. P. García Mayo (ed.): *Investigating Tasks in Formal Language Learning*. Multilingual Matters.

Larsen-Freeman, D. 2006. 'The emergence of complexity, fluency, and accuracy in the oral and written production of five Chinese learners of English,' *Applied Linguistics* 27: 590–619.

Loban, W. 1976. *Language Development: Kindergarten through Grade Twelve.* (Research Report No. 18). National Council of Teachers of English.

MacWhinney, B. 2000. *The CHILDES Project: Tools for Analyzing Talk. Volume II: The Database*, 3rd edn. Lawrence Erlbaum.

McCarthy, P. M. and **S. Jarvis.** 2007. '*Vocd*: A theoretical and empirical evaluation,' *Language Testing* 24: 459–88.

Meisel, J., H. Clahsen, and **M. Pienemann.** 1981. 'On determining developmental stages in natural second language acquisition,' *Studies in Second Language Acquisition* 3: 109–35.

Michel, M. C., F. Kuiken, and **I. Vedder.** 2007. 'The influence of complexity in monologic versus dialogic tasks in Dutch L2,' *International Review of Applied Linguistics in Language Teaching* 45: 241–59.

Mochizuki, N. and **L. Ortega.** 2008. 'Balancing communication and grammar in beginning-level foreign language classrooms: A study of guided planning and relativization,' *Language Teaching Research* 12: 11–37.

Monroe, J. H. 1975. 'Measuring and enhancing syntactic fluency in French,' *The French Review* 48: 1023–31.

Norris, J. M. 1996. *A validation study of the ACTFL Guidelines and the German speaking test*. Unpublished MA thesis, University of Hawai'i.

Norris, J. M. and **L. Ortega.** 2000. 'Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis,' *Language Learning* 50: 417–528.

Norris, J. M. and **L. Ortega.** 2003. 'Defining and measuring SLA' in C. Doughty and M. Long (eds): *Handbook of Second Language Acquisition*. Blackwell.

Norris, J. M. and **L. Ortega.** 2006. 'The value and practice of research synthesis for language learning and teaching' in J. M. Norris and L. Ortega (eds): *Synthesizing Research on Language Learning and Teaching*. John Benjamins.

Norris, J. M. and **P. Pfeiffer.** 2003. 'Exploring the use and usefulness of ACTFL guidelines oral proficiency ratings in college foreign language departments,' *Foreign Language Annals* 36: 572–81.

Oh, S. 2006. *Investigating the Relationship between Fluency Measures and Second Language Writing Placement Test Decisions*. Unpublished Master's Scholarly Paper. University of Hawai'i.

Ortega, L. 1995. *Planning and Second Language Oral Performance*. Unpublished MA thesis, University of Hawai'i.

Ortega, L. 1999. 'Planning and focus on form in L2 oral performance,' *Studies in Second Language Acquisition* 21: 109–48.

Ortega, L. 2003. 'Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing,' *Applied Linguistics* 24: 492–518.

Ortega, L. and **H. Byrnes.** 2008. 'Theorizing advancedness, setting up the longitudinal research agenda' in L. Ortega and H. Byrnes (eds): *The Longitudinal Study of Advanced L2 Capacities*. Routledge.

Perkins, K. 1980. 'Using objective methods of attained writing proficiency to discriminate among holistic evaluations,' *TESOL Quarterly* 14: 61–7.

Pienemann, M. 1998. *Language Processing and Second Language Development: Processability Theory*. John Benjamins.

Polio, C. 2001. 'Research methodology in second language writing research: The case of text-based studies' in T. Silva and P. K. Matsuda (eds): *On Second Language Writing*. Lawrence Erlbaum.

Révész, A. 2009. 'Task complexity, focus on form, and second language development,' *Studies in Second Language Acquisition* 31: 437–70.

Robinson, P. 2007. 'Task complexity, theory of mind, and intentional reasoning: Effects on L2 speech production, interaction, uptake and perceptions of task difficulty,' *International Review of Applied Linguistics* 45: 237–57.

Robinson, P. and **N. C. Ellis.** 2008. 'Conclusion: Cognitive linguistics, second language acquisition and instruction: Issues for research' in P. Robinson and N. C. Ellis (eds): *Handbook of Cognitive Linguistics and Second Language Acquisition*. Routledge.

Robinson, P. and **R. Gilabert.** (eds). 2007. *Task Complexity, the Cognition Hypothesis and Second Language Learning and Performance.* Special Issue of: *International Review of Applied Linguistics* 45/3.

Sangarun, J. 2005. 'The effects of focusing on meaning and form in strategic planning' in R. Ellis (ed.): *Planning and Task Performance in a Second Language*. John Benjamins.

Sato, C. 1990. *The Syntax of Conversation in Interlanguage Development*. Gunter Narr.

Scarborough, H. S. 1990. 'Index of productive syntax,' *Applied Psycholinguistics* 11: 1–22.

Scott, C. M. 1988. 'Spoken and written syntax' in M. Nippold (ed.): *Later Language Development: Ages Nine through Nineteen*. Little, Brown.

Skehan, P. 2003. 'Task based instruction,' *Language Teaching* 36: 1–14.

Skehan, P. and **P. Foster.** 2005. 'Strategic and on-line planning: The influence of surprise information and task time on second language performance' in R. Ellis (ed.): *Planning and Task Performance in a Second Language*. John Benjamins.

Storch, N. and **G. Wigglesworth.** 2007. 'Writing tasks: The effects of collaboration' in M. P. García Mayo (ed.): *Investigating Tasks in Formal Language Learning*. Multilingual Matters.

Tabachnick, B. G. and **L. S. Fidell.** 1996. *Using Multivariate Statistics*. HarperCollins.

Tavakoli, P. and **P. Foster.** 2008. 'Task design and second language performance: The effect of narrative type on learner output,' *Language Learning* 58: 439–73.

Tavakoli, P. and **P. Skehan.** 2005. 'Strategic planning, task structure, and performance testing' in R. Ellis (ed.): *Planning and Task Performance in a Second Language*. John Benjamins.

Tomasello, M. and **D. Stahl.** 2004. 'Sampling children's spontaneous speech: How much is enough?,' *Journal of Child Language* 31: 101–21.

Verspoor, M., W. Lowie, and **M. van Dijk.** 2008. 'Variability in second language development from a dynamic systems perspective,' *Modern Language Journal* 92: 214–31.

Wolfe-Quintero, K., S. Inagaki, and **H.-Y. Kim.** 1998. *Second Language Development in Writing: Measures of Fluency, Accuracy, and Complexity*. University of Hawai'i, Second Language Teaching and Curriculum Center.