Applied Linguistics 2017: 38/1: 90–111
doi:10.1093/applin/amv002 Advance Access published on 13 March 2015

The Development of Complexity, Accuracy, and Fluency in Second Language Performance: A Longitudinal Study

MARY LOU VERCELLOTTI

Department of English, Ball State University
E-mail: mlvercellott@bsu.edu

INTRODUCTION

Language, especially second language (L2), performance may be broken into linguistic subcomponents, including complexity, accuracy, and fluency (CAF). These subcomponents of language performance have been of increased interest in second language development (SLD). Generally, the objective in L2 learning is to master all three CAF subcomponents. In a cognitive framework of SLD, limited attentional resources inhibit learners from attending to all CAF components simultaneously. A focus on one CAF component may compromise a learner's performance in another CAF component, which has been called trade-off effects.

One goal of this study is to describe the development of oral language performance as measured by CAF. Understanding the effects of the processing demands of speaking in an L2 is of theoretical and pedagogical interest. Language programs may endorse tasks that promote development of each CAF component individually rather than expecting learners to be able to attend to every aspect of language performance during a real-time speaking task. SLD research has shown that certain tasks or task conditions can give learners opportunity to practice speaking with increased complexity, accuracy, or fluency (e.g. Yuan and Ellis 2003). Certainly, speakers may focus on one component, but must they? Another goal of this study is to explore whether trade-off effects are inevitable during L2 development by looking at the relationships between CAF components. These questions are best explored by considering observations nested within individuals.

In order to investigate the development of CAF, this study analyzed the linguistic performance from individual learners during multiple topic-based speeches, which were given over time in an intensive English program (IEP). This study's research design allows a better understanding of development than the previous research where the concluded trade-off effects, theoretically explained by limited attentional resources, were based on group means. This article first reviews research, from cognitive/information

processing and dynamic systems theory (DST) frameworks, which sheds light on the issues especially relevant to the development of CAF in L2 performance: (i) cognitive limitations in language performance, (ii) performance differences during different tasks or task conditions, and (iii) language performance over time. After describing the empirical results of the shape and speed of development and of the relationships among CAF over time, I argue that the results do not support the supposition of trade-off effects.

COGNITIVE LIMITATIONS IN LANGUAGE PERFORMANCE

Many researchers accept limitations in L2 performance from assumed competition in attentional resources. Simply put, focusing on one CAF component might result in a lower performance in one or both of the other components, that is, trade-off effects. From a cognitive framework, Skehan's Limited Capacity Hypothesis (1998) predicts a competitive relationship among CAF where adult learners emphasize meaning over form, which could potentially hinder further SLD. When learners do focus on form, according to Skehan, there is a secondary contrast between control of form (accuracy) and use of more advanced language (complexity). For Skehan, all language learners have these tensions during performance because of limited mental resources, specifically limited attentional capacity and working memory, accepting a single-source view of attention. Further, Skehan applies his hypothesis to development, outlining the need to apply pedagogical pressure in order for students to have balanced CAF development.

Even researchers who reject a single-source capacity limitation accept that trade-off effects may be found in language performance, explained by attentional control and interference (Robinson 2003). Robinson's Cognition Hypothesis (2011) expects tasks to promote either fluency or complexity *and* accuracy, which aligns with Skehan's primary trade-off but contrasts with the second. For instance, simple monologic tasks are likely to promote fluency (but not complexity or accuracy), while accuracy and complexity (but not fluency) are promoted during complex monologic tasks (Robinson 2011). Although the current study does not test Robinson's Cognition Hypothesis directly as it does not manipulate task complexity, it may shed light on the theoretical implications of the Cognition Hypothesis by supporting or refuting his expectations (or lack-there-of) of specific trade-off effects among the CAF components. Additionally, the findings from longitudinal studies such as this one are pertinent given Robinson's application of his hypothesis to language development and subsequent recommendations to curriculum design.

In DST, on the other hand, cognitive resources are limited but connected and possibly compensatory (de Bot 2008). All variables in the system are interrelated, so any and all changes will affect all the other parts of the system. Researchers who assume a DST or the similar complexity theory (Larsen-Freeman 2009) reject a cause-and-effect model of language learning (de Bot *et al*. 2007). Therefore, in this approach, specific trade-off effects may

be found, but they are not understood to have a causal, linear, or mutually exclusive relationship (de Bot *et al*. 2007). Limited resources do not always result in trade-off effects because 'connected growers' require fewer attentional resources than unconnected subsystems (Spoelman and Verspoor 2010). Consequently, a key to this theoretical approach is which subsystems have meaningful relationships (Verspoor *et al*. 2008). Although researchers using DST have predicted relationships within a developmental sequence of a single construct, the theory has not yet offered a developmental sequence across CAF constructs.

CAF COMPETITION DURING LANGUAGE PERFORMANCE

Although findings differ and sometimes contradict each other, researchers working in different theoretical frameworks have concluded that trade-off effects impact language performance. Some research has supported Skehan's primary competition between meaning and form. Grammatical complexity has been reported to increase at the expense of fluency (measured by the number of pauses) during an interview task (Bygate 2001). And, a trade-off between fluency and accuracy seems to be a particularly robust finding in the literature (Yuan and Ellis 2003; Michel *et al*. 2007; Ahmadian and Tavakoli 2011).

Some between-group research designs (Yuan and Ellis 2003; Ahmadian and Tavakoli 2011) have found students can have higher accuracy and complexity at the expense of fluency, supporting predictions in Robinson's Cognition Hypothesis but refuting Skehan's secondary contrast within form. Skehan and Foster (1997), however, report that complexity and accuracy seemed to have a competitive relationship during two of the three tasks in a study comparing the effect of planning. Likewise, Ferrari (2012) suggests a trade-off between complexity and accuracy.

Lexis is an additional necessary subcomponent of CAF (Skehan 2009b). Lexical retrieval is especially relevant to L2 oral fluency, where finding the right word might decrease fluency (Lennon 2000). Further, lexis has been reported to be in a competitive relationship and in a supportive relationship with both accuracy and grammatical complexity. Yuan and Ellis (2003) report a trade-off between lexical variety and accuracy in the oral narratives. In contrast, Robinson (1995) concludes that lexical variety and accuracy both increase in a more cognitively difficulty task. Skehan (2009a) also reports that for non-native speakers, lexical variety is positively correlated with accuracy but negatively correlated with grammatical complexity. Conversely, David *et al*. (2009) report lexical variety positively correlated with global grammatical complexity when aggregated across age groups.

A key question is whether trade-off effects, common in the literature, will be found when looking at individual performances. Often, conclusions of trade-off effects have been inferences from research using a cross-sectional design and based on group mean comparisons, which may not represent the

performances of the individuals. For instance, Yuan and Ellis (2003) studied the effect of planning on oral language performance and conclude that there was a trade-off effect between accuracy and fluency by comparing the means of different planning groups. The trade-off, however, was not found *within* each planning group. Similarly, Skehan and Foster (1997) report between-task group means as support for trade-off effects, even though trade-offs were not found *within* each task, which places the trade-off effects only at the study level, not at the group or individual level. Bygate (2001) describes trade-off effects when comparing performances during interviews and narratives, but there were no trade-offs *within* these tasks. And, Ishikawa (2007) reports no trade-off effects between complexity and accuracy in written texts within his task-complexity groups.

An emerging key explanatory variable in research purporting trade-off effects is the task or task instructions given to the groups during data collection. Different tasks or different instructions may encourage the learner to prioritize one component of the triad over the others (Ellis and Barkhuizen 2005). It seems that in many of the between-group designs which lead to conclusions of trade-off effects, the groups represent different CAF-focused performances. Although performances with trade-off effects can be induced, these findings do not establish an inevitable limitation of performances because of limited attentional capacity. It is unclear if the students *must* prioritize or *how* students will prioritize CAF without the effects of the differing demands of the task or task condition. Importantly, correlations between the CAF scores, particularly within-individual correlations which could illuminate whether individual students prioritized one construct over another, are not often considered or reported. In fact, a single study (Mizera 2006), which considered the relationship between the language performance scores, reports that accuracy and fluency were positively correlated. Crucially, cross-sectional designs with different CAF foci have limited applicability to theories of L2 development.

LANGUAGE DEVELOPMENT

Little work has been done to research language performance development (i.e. change), with many CAF studies focusing instead on performance status. CAF research from cognitive frameworks has not generally employed longitudinal designs, and few of the cross-sectional designs have used different proficiency groups to represent development. Researchers working from a DST framework have begun conducting longitudinal studies looking for relationships between CAF with some indications that a change in one CAF construct affects the development of another, but such research has tended to use written texts rather than spoken data (cf. Ferrari 2012; Polat and Kim 2014). For instance, Larsen-Freeman (2006) suggests that focusing on improving lexical variety may mean ignoring grammatical complexity.

For Higgs and Clifford (1982), accuracy development is compromised by a focus on lexical and fluency development. They propose that language learners

who are sufficiently proficient to communicate (with higher vocabulary and fluency) do not continue to develop grammatical accuracy because of proactive interference, in which learning to communicate interferes with the ability to subsequently learn how to communicate with accuracy. Competition over time within form (accuracy and complexity) has been investigated. Ahmadian (2011) concludes that the participants in his repeated-measures study attended to complexity at the expense of accuracy, based on an increase of Analysis of Speech unit (AS-unit) length but no increase in percentage of error-free clauses. In contrast, a study with written homework assignments found no meaningful relationship between accuracy and complexity in development (Spoelman and Verspoor 2010).

Individual differences, of course, can affect L2 performance, but this article only considers the impact of initial proficiency because participants entered the IEP at differing levels of proficiency. It is expected that higher proficiency students will have better initial scores, but initial proficiency was not held constant in the study for methodological reasons explained in the analysis section.

In summary, there are two main issues regarding SLD and trade-off effects. First, since most research from a cognitive framework has examined data collected at a single time-point (i.e. language performance status), it is unclear how CAF language performance changes over time (i.e. language performance development). More longitudinal research is needed in order to evaluate if each develops simultaneously or if specific patterns of growth limit development across CAF. Secondly, it is unclear if the trade-off effects often found in cross-sectional research (comparing group means, sometimes with different task conditions) will be found studying individual performances. This study used longitudinal oral performance data from English language learners from multiple language backgrounds (L1) to answer the following research questions.

RQ1: What are the developmental trajectories found in English language learners' CAF performances during monologues? I hypothesized that all measures would show improved performance over time, following previous research with written data (Larsen-Freeman 2006) but contrary to Higgs and Clifford's (1982) description of a 'terminal' spoken language profile.

RQ2: Does development reveal competitive or supportive CAF relationships? With conflicting results being reported from studies with different tasks and different measures, I made hypotheses based on the theoretical assumption of learners' limitations in attentional resources (Skehan 1998) or attentional control (Robinson 2011) during oral performances with little planning time. Hence, most CAF measures were expected to have a competitive relationship. However, I hypothesized that some CAF measures would be positively correlated. Specifically, lexical complexity and accuracy could be positively correlated because both measures can represent linguistic control,

perhaps native-like control (Skehan 2009b), and grammatical complexity and fluency could be positively correlated if learners have an attentional focus on expression (i.e. talkative but without regard to accuracy), which is possible within the Limited Capacity Hypothesis (Skehan 2003).

CAF DEVELOPMENT RESEARCH METHODOLOGY

This study was designed in a cognitive framework of language learning. This study is longitudinal, with multiple observations of the same individuals over several months.

Participants

Participants were students in an IEP in the USA in 2010. This study included participants with at least three speeches from the most common L1 backgrounds in the IEP (as part of a larger study which considered the influence L1 or cultural background): Arabic ($n = 43$), Chinese ($n = 16$), and Korean ($n = 7$). Among the 66 participants, 34 were male and 32 female, all young adults (range 18–35 years; M = 25.3 years; SD = 4.5). Upon enrollment, all were tested with a standardized test and two in-house assessments to be placed into instruction levels. The in-house listening placement test score was chosen as the best measurement of initial proficiency (and treated as the independent variable) because a Pearson correlation analysis indicated that the in-house listening test was most highly correlated with placement into instruction levels, $r = .838$ ($p < .001$). In other words, the scores on the listening test best predicted the human experts' evaluation of proficiency.[1] Moreover, explicit instruction in a sequenced IEP likely impacts language performance since instruction given at each level is substantially different with the levels having different instructional goals and texts, so addressing instruction level was important. The participants included students from two cohorts (i.e. enrolling subsequent academic semesters), but the cohorts were similar in age and initial proficiency scores, confirmed by a two-tailed t-test, $t(64) = -0.647$, $p = .520$ and $t(64) = -0.828$, $p = .411$, respectively. Most students in the IEP are simultaneously enrolled in speaking, listening, grammar, reading, and writing courses, which each meet for 50 minutes per day, four days a week. This IEP strives for a principally eclectic approach, employing communicative, task-based, and focus on form approaches.

Materials

This study included the coding and analysis of two-minute semi-spontaneous monologues[2] ($n = 294$) from the Recorded Speaking Activity in the IEP's curriculum. The assignments were roughly one month apart within semesters; observations across semesters were further apart. Not every participant remained in the IEP for three academic semesters. There were 4.45 (SD = 1.25)

observations per participant, with a minimum of 3 and maximum of 7, given over 3–10 months in an IEP. The number and topics differ from semester to semester and by instruction level.[3] The speeches were recorded during regular speaking class time in a language media lab on Apple Power Mac computers with software developed with Revolution Studio 2.6.1 (Shafer 2006). The task instructions asked the participants to speak on the given topic, such as 'describe your best friend . . .' (see Vercellotti 2012 for details). During the one-minute planning time, the students could not take notes or use reference materials. The classroom teachers graded the monologues with an analytic grading rubric that included elements of fluency, accuracy, grammatical and lexical complexity, which means that the students are not explicitly encouraged during data collection to prioritize one of the CAF components.

The speeches were transcribed using Praat (Boersma and Weenick 2007) by a native speaker of English, trained and experienced in transcribing non-native speech. The author checked each transcription and coded the data into clauses and AS-units which are defined base-units for oral language, following Foster *et al*. (2000). Utterances without copulas which were clearly completed utterances were coded as AS-units. Errors in syntax, morphology, and lexical choices were marked within the clauses. Whenever an utterance had a self-correction, only the final version was considered. Thus, an accurate self-correction could make that clause error-free, following Ellis and Barkhuizen (2005).

Analysis

The CAF of the language performance were quantified for each observation (i.e. speech). Complexity included grammatical complexity, calculated as mean length of AS-unit in words, and lexical variety, calculated as D using 'vocd' (McKee *et al*. 2000) based on word. D was chosen for its ability to reliably compare texts of different lengths, even relatively short texts (Durán *et al*. 2004), and it has been shown to be a useful measure for L2 data (Treffers-Daller 2009; Yu 2010). Accuracy was measured as percentage of error-free clauses, which is a general measure of accuracy. Ellis and Barkhuizen (2005) recommend a general measure of accuracy in most SLD research because specific measures may misrepresent learners' knowledge if learners avoid forms or constructions. Fluency was measured with mean length of pause (MLP) because pausing has been attributed to 'attentional preoccupation with micro-planning' (Schmidt 1992: 377), and this study was interested in possible trade-off effects from limited attentional resources. Following De Jong and Perfetti (2011), MLP was calculated as the average length of pause of at least 200 ms, including both silent and filled (e.g. 'uh') pauses. Combining silent and filled pauses is judicious because studies (e.g. de Jong *et al*. 2015) have shown that L2 speakers tend to use either filled or silent pauses, depending on individual (L1 and L2) speaking style.

To answer the first research question about the developmental trajectory of each CAF measure, the data were analyzed using Hierarchical Linear and

Non-linear Modeling (HLM). HLM has similarities to linear regression and analysis of covariance (ANCOVA). HLM can model longitudinal data, capturing change in performance (Singer and Willett 2003). Longitudinal data, with multiple observations from each individual, have increased analytic complexity, which is both a blessing and a curse. First, with multiple observations from each participant, the intervals between the observations differ due to weekly and semester schedules. Secondly, the number of observations per participant varies because of attrition. Thirdly, an individual's observations are expected to be more correlated than observations from different individuals. Statistical analysis methods based on the general linear model (e.g. ANCOVA) are not recommended for these complications; a method with mixed-modeling or multi-level modeling is required. (See Cunnings 2012 for an overview of the value of these models in SLD research.) Crucially, HLM addresses these concerns. HLM considers the distance between observations in the model. HLM allows each participant's trajectory to have a unique number of observations. Since incomplete data sets are not excluded, the study better represents the population. And, HLM analyzes observations nested within individuals.

HLM models include a coefficient for the intercept (often the initial observation), a coefficient for the slope, and potentially coefficients for higher-order terms (i.e. factors that affect multiple observations). When there is variation between individuals, random effects must be included in the model. For instance, when some individuals have higher initial scores, the model must allow for differing initial scores. The variance component listed under random effects is larger with more variation. With the well-known challenge of variability in SLD, covariates can be added to the model to explain differences in the scores at the initial observation and/or differences in the slope. Overall, an HLM analysis allows an exploration of individual trajectories over time, rather than comparing group averages at single time points.

The data were fitted using full-maximum likelihood HLM, using HLM6 (Raudenbush *et al*. 2004). For each CAF measure, time was adjusted by approximately one month (0.833 fraction of the year from the start of the semester) in the growth model so that the intercept was approximately at the start of data collection. For each, a chi-square test was performed to compare linear and non-linear growth models, but only the results of the final model, not the statistics evaluating competing models, are reported. Initial proficiency was tested as an independent variable in each model because the participants varied in proficiency upon entrance in the IEP.[4] Initial proficiency was not held constant because the research question considers growth over time in an instructed context, rather than growth from a specific point. Moreover, there is always differing proficiency among students, even within a level (e.g. intermediate), but an interval variable can capture this variation. The chosen initial proficiency measure was entered as a centered, continuous[5] variable. Table 1 lists the independent and dependent variables in this study.[6] Lastly, other CAF measures, which are internal time-varying predictors[7] (Singer and Willett 2003), were tested in each model to check if

Table 1: Summary of independent and dependent variables

Independent variables		Dependent variables (also possible internal *time-varying* predictors)
Time-invariant	Time-varying	
Initial proficiency	Topic	Complexity score (length of AS-unit)
	Clause length	Lexical variety score (D score)
		Accuracy score (percentage of error-free clauses)
		Fluency score (mean length of pause)

these scores partially explain any variation in the model. In other words, the other CAF measures were considered potential predictors in order to test for trade-off effects. Only the results of the final model are fully reported. Partial results of some interim models which address possible trade-off effects are noted because when other dependent measures are (or are not) significant in a model, the HLM results begin to answer the second research question regarding possible competitive or supportive relationships between the measures.

In order to more fully answer the second research question of whether development reveals competitive or supportive relationships among the CAF measures, the pooled within-individual correlations were calculated on the full data set using EQS. Within-individual correlations are valuable when considering possible trade-off effects within language performances, while between-individual correlations (such as typical Pearson correlations) describe group performances.

CAF DEVELOPMENT RESULTS

I hypothesized growth in all measures, and that hypothesis was confirmed. The best-fitting conditioned (including initial proficiency and any time-varying predictors) models are described below.

Complexity

Results of the best-fitting linear growth model included initial proficiency in the initial score (intercept) and the covariate fluency (MLP) in the slope model. Fluency is a plausible predictor in a cognitive framework because extra online planning time (i.e. longer pauses) might be needed to increase grammatical complexity (e.g. Yuan and Ellis 2003). The expected complexity score for average-proficiency students was estimated to be 11.584 words at one month (shown as the coefficient for the intercept in Table 2). The proficiency coefficient suggests that for every unit increase in centered initial proficiency, there was a corresponding increase in complexity scores by 0.200 words. The fluency

Table 2: Conditioned linear growth model of grammatical complexity

Fixed effects	Coefficient	SE	*t*	*p*
Model for initial status, π_0				
Mean length of AS-unit (words), β_{00}	11.584	0.488	23.738	<.001
Initial proficiency, β_{01}	0.200	0.039	5.080	<.001
Model for growth rate, π_1				
Mean growth rate, β_{10}	6.564	1.111	5.909	<.001
Mean length of pause, β_{20}	−1.683	0.446	−3.772	<.001
Random effects	**Variance component**	***df***	X^2	***p***
Initial status, r_0	0.478	64	72.256	.224
Change rate, r_1	20.833	65	97.266	.006
Mean length of pause slope, r_2	0.431	65	74.034	.207
Level-1 error, e	4.678			

coefficient suggests that for every increase in MLP, the length of the AS-unit decreased by 1.683 words. Plainly, *shorter* AS-units (*lower* grammatical complexity) are associated with *longer* pauses (*lower* fluency), which is contrary to a trade-off effect. The mean linear growth rate for participants was 6.564 words over the course of a year in the IEP.

Students did not vary significantly in their complexity scores at one month ($\chi^2_{64} = 72.256$, $p = .224$) after controlling for proficiency. This means that the remaining variance in initial scores was small enough for the differences to be considered explained after considering the participant's initial proficiency. The student scores varied significantly in growth rate at one month ($\chi^2_{64} = 97.266$, $p = .006$) even after controlling for fluency. In other words, the participants' growth trajectories differed (significant variance remained), but a more parsimonious, better-fitting model was not found using the variables considered in this study.

When lexical variety was tested to the model, the results showed an increase in lexical variety scores was associated with *higher* grammatical complexity, that is, speeches with higher lexical variety scores had longer AS-units. As stated, the simpler model better fit the data, but this finding is reported because the literature review suggested a possible trade-off between grammatical complexity and lexical variety.

In summary, the grammatical complexity results indicated that higher initial proficiency scores corresponded with higher initial scores, as would be expected. The proportion of initial score variance explained[8] by initial proficiency was calculated to be 80.0%. The best-fitting model was a linear growth

Table 3: Conditioned quadratic growth model of lexical variety (D)

Fixed effects	Coefficient	SE	*t*	*p*
Model for initial status, π_0				
Mean lexical variety (D) score, β_{00}	53.390	1.39	38.476	<.001
Initial proficiency, β_{01}	1.009	0.241	4.184	<.001
Model for growth rate, π_1				
Mean growth rate, β_{20}	−26.240	10.901	−2.407	.017
Mean acceleration rate, β_{30}	61.674	20.077	3.072	.002
Random effects	**Variance component**	***df***	X^2	*P*
Initial status, r_0	38.680	64	131.513	<.001
Level-1 error, e	173.383			

trajectory, and lower fluency (higher MLP) corresponded with lower complexity scores. Additionally, an increase in lexical variety scores was associated with higher grammatical complexity scores.

For lexical variety, a non-linear trajectory better fit the data, so the data were fitted with a quadratic growth model. Since the variance component of the non-linear trajectory was not significant (i.e. the differences in the growth trajectories between individuals were not large enough or consistent enough to try to explain the variance) it was constrained to zero following standard HLM procedure.

The results of the conditioned quadratic growth model (Table 3) found initial proficiency as a predictor of initial scores. The expected lexical variety score (D) for an average student at one month was estimated to be 53.390. The proficiency coefficient suggests that for every point increase in centered initial proficiency, there was a corresponding 1.009 increase in lexical variety scores. The mean linear change rate at one month was estimated to be −26.240, and the mean acceleration was estimated to be 61.674, which means the participants' scores showed a decrease followed by a steeper increase in lexical variety scores over time. The initial lexical variety scores varied significantly at one month ($\chi^2_{64} = 131.513$, $p < .001$) after controlling for initial proficiency.

Accuracy was tested in the model to explore RQ2, even though accuracy is not theoretically expected to predict lexical variety scores, because all covariates were tested in all models. The results suggested that for every point increase in accuracy, there was a corresponding 15.996 increase in lexical variety scores.

In summary, the results for lexical variety indicate that the growth trajectory was non-linear (with a dip and then a steeper increase) and that participants with higher initial proficiency had higher initial lexical variety scores. Overall,

initial proficiency explained 46.0% of the variance in initial scores in lexical variety. Additionally, increased accuracy scores were associated with higher lexical variety scores.

Accuracy

Results of the best-fitting linear growth model for accuracy found that initial proficiency and clause length were significant predictors. Since the variance component for change rate was not significant, it was constrained to zero. Results of the linear growth model with time-varying covariate of clause length (Table 4) specified that for average students with average clause-length scores, the expected clause accuracy score at one month was estimated to be 0.852 (85.2% of clauses were error-free). For every one unit increase in centered initial proficiency, there was a 0.012 (1.2%) increase in percentage of error-free clauses. At a certain time point, a one-word increase in clause-length scores decreased the clause accuracy scores by 0.046 (4.6%). The mean linear growth rate for all students was estimated to be 0.088 (8.8%) while controlling for clause-length scores. After controlling for clause length, students no longer varied significantly in their accuracy scores at one month in the IEP ($\chi^2_{64} = 60.406$, $p > .500$) or in the relationship between clause accuracy and clause-length scores ($\chi^2_{65} = 65.898$, $p = .446$).

In summary, the results indicate that participants with higher initial proficiency scores had higher initial accuracy scores. The proportion of initial score

Table 4: Conditioned linear growth model of accuracy with covariate clause length

Fixed effects	Coefficient	SE	*t*	*p*
Model for initial status, π_0				
Mean accuracy score, β_{00}	0.852	0.045	18.990	<.001
Initial proficiency, β_{01}	0.012	0.002	4.945	<.001
Model for growth rate, π_1				
Mean growth rate, β_{10}	0.088	0.038	2.343	.020
Mean clause length slope, β_{20}	−0.046	0.008	−5.896	<.001
Random effects	**Variance component**	***df***	X^2	***p***
Initial status, r_0	0.0007	64	60.406	>.500
Clause length Slope, r_2	0.0004	65	65.898	.446
Level-1 error, e	0.012			

variance explained by initial proficiency was 8.6%. Longer clauses were less likely to be error-free. The change trajectory showed linear growth in accuracy over time in the IEP.

Fluency

The best-fitting model included initial proficiency and lexical variety scores as predictors of fluency scores (Table 5). For average students with average lexical variety scores at one month in the IEP, the expected MLP score was estimated to be 1.163 s. The coefficient for proficiency suggests that for every increase of centered initial proficiency, there was a corresponding decrease in MLP scores (higher fluency) by 0.017 s. At a certain time point, one point increase in lexical variety score further decreased MLP scores (higher fluency) by 0.003 s. The mean linear growth rate was estimated to be −0.609 s (improved fluency) while controlling for lexical variety. Students still varied significantly in their MLP scores at 1 month ($\chi^2_{35} = 286.097$, $p < .001$), change rate ($\chi^2_{36} = 152.862$, $p < .001$), and in the relationship between MLP and lexical variety at a given time point ($\chi_{36} = 188.769$, $p = .001$).

In summary, the results indicate that initial proficiency and lexical variety scores predicted fluency scores. Higher initial proficiency corresponded with slightly shorter pauses. The proportion of initial score variance explained by initial proficiency was found to be 16.0%. Higher lexical variety scores corresponded with shorter pauses (*better* fluency). The change trajectory showed

Table 5: Conditioned linear model of growth in mean length of pause with lexical variety

Fixed effects	Coefficient	SE	*t*	*p*
Model for initial status, π_0				
Mean length of pause, β_{00}	1.163	0.104	11.231	<.001
Initial proficiency, β_{01}	−0.017	0.007	−2.28	.026
Model for growth rate, π_1				
Mean growth rate, β_{10}	−0.609	0.125	−4.87	<.001
Mean lexical growth rate, β_{20}	−0.003	0.002	−2.01	.048
Random effects	**Variance component**	***df***	X^2	***p***
Initial status, r_0	0.473	35	286.097	<.001
Change rate, r_1	0.506	36	152.862	<.001
Lexical variety slope, r_2	0.0008	36	188.769	<.001
Level-1 error, e	0.030			

Table 6: Within-individual correlations for CAF measures

	Accuracy	Complexity	Lexical	Fluency
Accuracy (percentage error-free clauses)	–			
Grammatical complexity (length of AS-unit)	0.139**	–		
Lexical variety (D) score	0.125*	0.201**	–	
Fluency (MLP with polarity reversed)	0.108*	0.363**	0.231**	–

Note: *Significant at $p < .05$ level; **significant at $p < .01$ level.

improved fluency over time. There was still significant variation in the initial scores and in the growth trajectories, which could not be explained by the variables considered in this study.

Correlation analysis

The second research question concerned the relationships between constructs. The within-individual correlation analysis (Table 6) found mostly modest, yet significant, positive relationships between the CAF measures, rather than competitive relationships, confirming the HLM results. In order to more easily interpret the correlations table, the polarity of the fluency (measured as MLP) is reversed so that positive scores reflect better language performance.

The within-individual correlations showed that accuracy and grammatical complexity were positively correlated ($r = .139$), indicating that higher complexity scores (length of the AS-unit), correlated with higher accuracy scores. Likewise, lexical variety and accuracy had a significant positive correlation ($r = .125$), and grammatical complexity and lexical variety were positively correlated ($r = .201$). The accuracy and fluency correlation ($r = .108$) indicated that higher accuracy scores correlated with better fluency scores. The grammatical complexity and fluency scores ($r = .363$) and the lexical variety and fluency scores ($r = .231$) were also positively correlated. Most of the correlations are considered weak, which is expected because the CAF measures capture different aspects of language performance. All correlations are statistically significant, and importantly, the polarity of all relationships is contrary to trade-off effects.

DISCUSSION

Grammatical complexity, accuracy, and fluency had linear change trajectories, each showing improvement. Lexical variety had a non-linear trajectory, showing a slight decline and followed by steeper increase over time.

Expectedly, initial proficiency was a predictor of initial scores for each measure; higher initial proficiency predicted better initial performance. Topic was not found to be a statistically significant predictor. This finding is likely from

the variability in the suspected topic effects (i.e. some topics encouraged lexical variety, some did not; some topics displayed large variation in scores, some did not). The topics (chosen for pedagogical reasons) given to the students may have influenced the lexical variety scores in unplanned ways, which has been found in other research (e.g. Yu 2010), but the models could not be improved by including inconsistent topic effects.

Negative influence between some CAF constructs was expected but not found. Although an increase in length of unit corresponded to a decrease in accuracy (i.e. longer clauses were less likely to be error-free), this finding is understood as a result of the calculation of accuracy as proportion of error-free units and does not necessarily support the notion trade-offs. Ferrari's (2012) conclusion of a trade-off effect between accuracy and complexity was based on similar measures without controlling for the increasing length of utterances over time. Controlling for clause length with accuracy measures has been discussed in the field (e.g. Skehan and Foster 2005).

Interestingly, lexical variety was a relevant predictor in the fluency model, in that students with higher lexical variety scores had higher fluency, and lexical variety scores were correlated with better fluency, contrary to a trade-off effect. It seems that retrieval of varied lexical items did not require longer pauses. Both the HLM and the within-individual correlational results, which are contrary to the expectation of the cost of lexical retrieval, may support a model (e.g. MacWhinney 2001) where lexical retrieval occurs before the construction of the syntactic frame (rather than during formulation). Further, the additional HLM model testing indicated that higher lexical variety was related to accuracy and grammatical complexity as well. Considering the HLM results and the within-individual correlation results, lexical variety did not hinder but promoted fluency, accuracy, and grammatical complexity, which may endorse vocabulary instruction in IEPs. Alternatively, this finding may indicate that lexical variety serves as another measure of general proficiency.

Admittedly, the different scales of measurements make comparisons across constructs difficult to interpret, and some coefficients were small. With continuous variables, however, the effects can accumulate. Most importantly, the polarity of all coefficients in the HLM models challenge the expected trade-off effects. Likewise, the within-individual correlation analysis found all of the CAF constructs were positively correlated. Considering that all measures showed growth over time and the within-individual correlations were positive, an increase in one CAF measure was correlated with an increase in the others. Although it might be presumed that learners cannot be accurate and fluent, these results showed that students did not sacrifice fluency for increased accuracy, echoing the correlations found in Mizera (2006).

This study has two major theoretical implications: these ESL students showed similar growth trajectories and there was a lack of trade-off effects in these topic-based speeches. First, this study showed generally shared paths to development in students learning in the same IEP even though

there was intra-individual variation and some measures (accuracy and lexical variety) did not have significant between-individual variation in trajectories to investigate separate paths of development. This finding is contrary to Larsen-Freeman's (2006) longitudinal study which suggested that there may be 'preferred paths'. Larsen-Freeman reports findings based on five learners' written texts on the same topic using different measures than the current study, and the group mean increased over time for each measure. The differences in mean scores were not statistically significant, but from DST approach, intra-individual variation is of interest. In the plot illustrating 'distinctive' paths between grammatical and lexical complexity over time, the distinctly different learner had a higher lexical score than the others throughout the study, which means that the distinctiveness of that learner was not a matter of development. Further, no statistical analysis was completed to determine if there was any significant individual CAF development. A look at the individual scores reveals that three participants had some growth in both grammatical and lexical complexity, one participant had almost no change across all four observations in either measure, and one participant had no change for one measure and very little change in the other. Hence, although her paper serves a pioneering role in CAF development (rather than status) there is really not much support for separate paths over development in her data.

In contrast, the current study found that the within-individual correlations were positive and highly significant, meaning that students did not focus their development on one CAF construct (e.g. grammatical complexity or lexical complexity) at the expense of another. The constructs grew together. There was remaining variation in the initial score of lexical variety, the initial score of the fluency measure and the slope of the fluency measure, but the available predictors in the study, even time-varying predictors (i.e. scores on the other measures), did not build a more parsimonious model. Consequently, even though some variation remains unexplained, the findings did not support a separate path explanation.

The results were also contrary to Higgs and Clifford's (1982: 73) suggestion from 'experiential but consistent data' that communication success (i.e. getting your idea across with vocabulary and fluent speech) inhibits the need to produce grammatically accurate language. Higgs and Clifford's hypothesis is directional, that is, sufficient vocabulary and fluency inhibits continued growth in accuracy. The current study found that accuracy was not negatively affected by higher vocabulary or fluency scores. The accuracy scores showed growth over time without evidence for a plateau in development, and importantly, the HLM testing found that neither fluency nor lexical variety scores were significant in the accuracy model. Additionally, the within-individual showed significant positive correlations among CAF. Using these quantitative analyses, high fluency, and/or high lexical variety did not negatively impact accuracy. These trajectory patterns, therefore, indicate that individual differences in performance are more likely a result of developmental lag not developmental deficit at this stage of SLD.

Secondly, despite the intuitive appeal of the inevitability of trade-off effects, these longitudinal data suggest that individual development did not show CAF trade-off effects, not even the oft-cited fluency-accuracy trade-off. Although the operationalization of the CAF constructs (i.e. the chosen measurements) possibly affected the conclusions, I propose that data analysis (group means vs. individual growth curves), research design (cross-sectional vs. longitudinal), and task design (different CAF foci vs. no explicit CAF focus) explain this major difference in findings. First, previous research, for the most part, reports on aggregate data. However, multi-level analysis and within-individual correlations are much more robust for considering trade-off effects within individual performances than comparing group means. Actual trade-off effects should be detected at the individual level because trade-off effects are hypothesized to be exerted within the individual. Moreover, aggregating data inflates correlations (Ostroff 1993), which could lead to apparently significant correlations that would not be found in non-aggregated data.

Any artificial inflation of results from the statistical method, however, does not elucidate the findings completely contrary to trade-off effects. A skeptical reader might question whether the development of all measures drove the positive correlations. Subsequently, separate between-individual correlations were run on the observations at the first four time points, which would be similar to the correlation analysis employed in studies looking at performance status at a single time (e.g. Mizera 2006). Although these correlations do not always reach significance at each time point, the results follow the same pattern as the longitudinal analysis (i.e. connected growers), which are again contrary to previously reported trade-off effects.

This study showed that CAF development did not come at the expense of another construct, despite variations found in individual performances. This research highlights how mean group performance (aggregated data) may not necessarily reflect individual performances, even though they are often used to reflect individuals' performance. Critically, many of the cross-sectional studies citing trade-off effects based on group means, the groups actually performed different 'tasks' or the same task under different task instructions. Ishikawa (2007) made a similar argument, stating that different task demands may direct learners' attention to one CAF construct. Inferences of trade-off effects, especially with cross-sectional designs, should be made conservatively, and only when the improvement in one construct comes *at the expense of* another. The more sophisticated multi-level modeling (HLM) and the within-individual correlations found no support for a trade-off hypothesis, contra Skehan's meaning vs. form (or complexity vs. control). Likewise, these results only partially support Robinson's (2011) Cognition Hypothesis, in that accuracy and complexity can both be attended to. But, Robinson still expects a competition between fluency and form (whether the task is considered simple or complex) which was not found here.

Since many studies with conclusions of trade-offs did not involve different proficiency levels (but rather different tasks or task demands), the groups do not

represent development. Despite this, both Skehan and Robinson connect their theories to curriculum design, citing research described in the literature review as support. Robinson (2006) explicitly states that task complexity affects learning not just performance, and that the predictions made by the Cognition Hypothesis is applicable to L2 development. Likewise, Skehan (1998: 288), clearly influenced by the near-inevitability of trade-off effects, argues that curriculum must systematically manipulate learner's attention 'to focus on particular aspects of language performance'. The current study's findings challenge this recommendation. Balanced development (the ultimate goal of SLD) might arise without the systematic manipulation of attentional resources in a task-based curriculum since the language elicited from a single speaking task showed development of all CAF constructs. The current findings, which illuminate the process of CAF development more directly than previous research, may be more confidently applied to the theories of curriculum design.

These findings may elucidate the nature of SLD and the mechanisms behind language production. For instance, these results may challenge a single-source attentional pool. On the other hand, they may indicate that as L2 learning becomes proceduralized, less attentional resources are required, allowing learners to attend to multiple CAF constructs. The findings are congruent with DST in that at least some constructs may be considered 'connected growers'. DST remains underspecified, however, if it only states all CAF constructs are connected. Alternatively, the positive correlations between constructs may be driven by a higher level factor. For instance, Dewaele and Furnham (1999) state that extraverts are better at parallel-processing. Considering the demands of L2 performance, being able to process in parallel could allow a speaker to maintain fluency while complexifying and monitoring her speech.

CONCLUSION

Using multi-level analysis with longitudinal data allowed a detailed interpretation of SLD. The learners had gains in all CAF constructs, which supports a view of CAF as connected growers. Importantly, the generally accepted trade-off effects were not evident these topic-centered monologues. In fact, the within-individual correlation showed significant positive correlations between each CAF measure. Based on multiple observations from 66 participants, the HLM analyses did not support bifurcated paths of development by CAF. Initial proficiency affected initial CAF scores, but any remaining variation in scores was not the focus of this study and left for future research.

Given the intricacies of longitudinal data, the data analyses used in this study (HLM and within-individual correlations) better explore multiple observations of individual participants and the interconnectedness of language development. Considering the variability in linguistic performances, these statistical methods, rather new to the field of applied linguistics, can lead to enhanced investigation of the complexity of SLD. The results of studies such as this one impact language learning theories, suggesting an interconnected or

holistic view of language proficiency and prompting more research on understanding attentional resources and automatization within a connected language system. Moreover, such findings have pedagogical implications on curriculum, for proficiency testing, and for program evaluation. A number of limitations must be noted. This study was limited to students at a single IEP and to the proficiency range of these students, and the overall findings could be affected by this IEP's curriculum. The research should be replicated in other IEPs and with different populations, specifically populations with other language backgrounds and with students studying English as a foreign language. Even though the observations spanned three academic semesters, a longer study or a study of participants at a different stage of SLD may find different trajectories. In addition, although HLM can test for non-linear models, forcing the data into any particular form may be objectionable within DST. The conclusions are limited to the language performance elicited by open-ended topic prompts because topic avoidance (Tarone 1980) may be a viable strategy with this prompt type and its effect (if any) was not measured. Although the results presented are representative of the larger study employing multiple measures of CAF reported in Vercellotti (2012), other measures could give a more complete picture or even a different picture. Also, more research is needed to explain the remaining variance. For instance, independent variables, either time-invariant (e.g. assertiveness as suggested by Ockey 2011) or time-varying (e.g. interest in or familiarity of the topic) might help explain differences in fluency scores. The inclusion of these possible individual differences affecting the perceived task difficulty (Robinson 2011) might explain some inter- and intra-individual variation.

ACKNOWLEDGEMENTS

This research was funded by National Science Foundation, Grant Number SBE-0836012, to the Pittsburgh Science of Learning Center (PSLC, http://www.learnlab.org). This article is based on my dissertation research; many thanks to my committee, Nel de Jong, Kevin Kim, Claude Mauk, Alan Juffs, Dawn E. McCormick, and Yas Shirai, for their superb guidance, and Anthony Bohan for his help with calculations. I also thank Liz Riddle, Elmar Hashimov, Lourdes Ortega, Anna Mauranen, and the journal's anonymous reviewers for their encouraging words and their insightful, useful comments. The data can be found at http://talkbank.org/browser/index.php?url=SLABank/English/Vercellotti/.

NOTES

1 The Michigan Test of English Language Proficiency scores of students placed in the low-intermediate (range 32–77; M = 48.4; SD = 11.4) and high-intermediate instruction level (range 37–80; M = 60.2; SD = 11.5) had substantial overlap and correlated less closely (r = .642).

2 Monologues allow for more complex utterances compared with dialogues (Ferrari 2012). The pragmatics of dialogues may interfere with the study of relationships among CAF.

3 Although multiple topics increase variability, this variability separates the sequence of topics from development

(i.e. if all had the same first topic and the same sequence of topics, any change in scores could be driven by unknown and unplanned topic effects).

4 Initial proficiency was also tested when there was significant variation in the random effects in the slope, but it was not a significant predictor of slope (i.e. change rate).

5 In models with a centered continuous variable as predictor, the coefficient indicates the difference in score per unit away from the mean, similar to regression analysis.

6 Clause length, clauses/AS-unit, phonation time ratio, mean length of run, and proportion of error-free AS-units were considered dependent measures in a larger study; those results (consistent with conclusions here) can be found in Vercellotti (2012).

7 A time-varying predictor is similar to a covariant in ANCOVA.

8 Explained variance is often reported with HLM results, in lieu of effect size.

REFERENCES

Ahmadian, M. J. 2011. 'The effect of 'massed' task repetitions on complexity, accuracy and fluency: Does it transfer to a new task?,' *Language Learning Journal* 39/3: 1–12

Ahmadian, M. J. and **M. Tavakoli.** 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.

Boersma, P. and **D. Weenink.** 2007. 'Praat (Version 4.5.01) [Computer software],' available at http://www.fon.hum.uva.nl/praat/.

Bygate, M. 2001. 'Effects of task repetition on the structure and control of oral language,' in M. Bygate, P. Skehan, and M. Swain (eds): *Researching Pedagogic Tasks: Second Language Learning, Teaching and Testing*. Pearson Education Limited.

Cunnings, I. 2012. 'An overview of mixed-effects statistical models for second language researchers,' *Second Language Research* 28/3: 368–82.

David, A., F. Myles, V. Rogers, and **S. Rule.** 2009. 'Lexical development in instructed L2 learners of French: is there a relationship with morphosyntactic development' in B. Richards, D. Malvern, P. Meara, J. Milton, and J. Treffers-Daller (eds).

de Bot, K. 2008. 'Introduction: second language development as a dynamic process,' *Modern Language Journal* 92/2: 166–78.

de Bot, K., W. Lowie, and **M. Verspoor.** 2007. 'A dynamic systems theory approach to second language acquisition,' *Bilingualism: Language and Cognition* 10/1: 7–21.

De Jong, N. and **C. A. Perfetti.** 2011. 'Fluency training in the ESL classroom: an experimental study of fluency development and proceduralization,' *Language Learning* 61/2: 1–36.

De Jong, N. H., R. Groenhout, R. Schooen, and **J. H. Hulstijn.** 2015. 'Second language fluency: speaking style or proficiency? Correcting measures of second language fluency for first language behavior,' *Applied Psycholinguistics* 36/2: 223–43.

Dewaele, J. M. and **A. Furnham.** 1999. 'Extraversion: the unloved variable in applied linguistic research,' *Language Learning* 49/13: 509–44.

Durán, P., D. Malvern, B. Richards, and **N. Chipere.** 2004. 'Developmental trends in lexical diversity,' *Applied Linguistics* 25/2: 220–42.

Ellis, R. and **G. Barkhuizen.** 2005. *Analysing Learner Language*. Oxford University Press.

Ferrari, S. 2012. 'A longitudinal study of complexity, accuracy and fluency variation in second language development' in A. Housen, F. Kuiken, and I. Vedder (eds): *Dimensions of L2 Performance and Proficiency*. John Benjamins.

Foster, P., A. Tonkyn, and **G. Wigglesworth.** 2000. 'Measuring spoken language: a unit for all reasons,' *Applied Linguistics* 21/3: 354–75.

Higgs, T. and **R. Clifford.** 1982. 'The push toward communication' in T. Higgs (ed.): *Curriculum Competence and the Foreign Language Teacher*. National Textbook Company.

Ishikawa, T. 2007. 'The effect of manipulating task complexity along the (±Here-and-Now) Dimension on L2 written discourse' 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* 27/14: 590–619.

Larsen-Freeman, D. 2009. 'Adjusting expectations: the study of complexity, accuracy, and fluency in second language acquisition,' *Applied Linguistics* 30/4: 579–89.

Lennon, P. 2000. 'The lexical element in spoken second language fluency' in H. Riggenbach (ed.): *Perspectives on Fluency*. University of Michigan Press.

MacWhinney, B. 2001. 'The competition model: the input, the context, and the brain' in P. Robinson (ed.): *Cognition and Second Language Instruction*. Cambridge University Press.

McKee, G., D. Malvern, and **B. Richards.** 2000. 'Measuring vocabulary diversity using dedicated software,' *Literary and Linguistic Computing* 15: 323–37.

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* 45/13: 241–59.

Mizera, G. J. 2006. 'Working memory and L2 fluency,' Ph. D. dissertation, University of Pittsburgh.

Ockey, G. 2011. 'Self-consciousness and assertiveness as explanatory variables of L2 oral ability: a latent variable approach,' *Language Learning* 61/3: 968–89.

Ostroff, C. 1993. 'Comparing correlations based on individual-level and aggregated data,' *Journal of Applied Psychology* 78/4: 569–82.

Polat, B. and **Y. Kim.** 2014. 'Dynamics of complexity and accuracy: A longitudinal case study of advanced untutored development,' *Applied Linguistics* 35/2: 184–207.

Raudenbush, S. W., A. S. Bryk, Y. Cheong, and **R. T. Congdon.** 2004. *HLM 6 Hierarchical Linear Modeling*. Scientific Software International.

Richards B., M. Daller, D.D. Malvern, P. Meara, J. Milton, and **J. Treffers-Daller** (eds). 2009. *Vocabulary Studies in First and Second Language Acquisition: The Interface Between Theory and Application*. Palgrave Macmillian.

Robinson, P. 1995. 'Task complexity and second language narrative discourse,' *Language Learning* 45/1: 99–140.

Robinson, P. 2003. 'Attention and memory during SLA' in C. J. Doughty and M. H. Long (eds): *The Handbook of Second Language Acquisition*. Blackwell Publishing.

Robinson, P. 2006. 'The cognition hypothesis, task design, and adult task-based language learning,' *Second Language Studies* 21/2: 45–105.

Robinson, P. 2011. 'Second language task complexity, the cognition hypothesis, language learning and performance' in P. Robinson (ed.): *Second Language Task Complexity: Researching the Cognition Hypothesis of Language Learning and Performance*. John Benjamins.

Schmidt, R. 1992. 'Psychological mechanisms underlining second language fluency,' *Studies in Second Language Acquisition* 14/14: 357–85.

Shafer, D. 2006. *Revolution: Software at the Speed of Thought*. Shafer Media.

Singer, J. D. and **J. B. Willett.** 2003. *Applied Longitudinal Data Analysis*. Oxford University Press.

Skehan, P. 1998. *A Cognitive Approach to Language Learning*. Oxford University Press.

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

Skehan, P. 2009a. 'Lexical performance by native and non-native speakers on language-learning tasks' in B. Richards, D. Malvern, P. Meara, J. Milton, and J. Treffers-Daller (eds).

Skehan, P. 2009b. 'Modelling second language performance: integrating complexity, accuracy, fluency, lexis,' *Applied Linguistics* 30/4: 1–23.

Skehan, P. and **P. Foster.** 1997. 'Task type and task processing conditions as influences on foreign language performance,' *Language Teaching Research* 1/3: 185–211.

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.

Spoelman, M. and **M. Verspoor.** 2010. 'Dynamic patterns in development of accuracy and complexity: a longitudinal case study in

the acquisition of Finnish,' *Applied Linguistics* 31/4: 532–53.

Tarone, E. 1980. 'Communication strategies, foreigner talk, and repair in interlanguage,' *Language Learning* 30/2: 417–31.

Treffers-Daller, J. 2009. 'Language dominance and lexical diversity: how bilinguals and L2 learners differ in their knowledge and use of French lexical and functional items' in B. Richards, D. Malvern, P. Meara, J. Milton, and J. Treffers-Daller (eds).

Vercellotti, M. L. 2012. 'Complexity, accuracy, and fluency as properties of language performance: The development of the multiple subsystems over time and in relation to each other,' Ph. D. dissertation. University of Pittsburgh.

Verspoor, M., W. Lowie, and **M. Van Dijk.** 2008. 'Variability in second language development from a dynamic systems perspective,' *Modern Language Journal* 92/2: 214–31.

Yu, G. 2010. 'Lexical diversity in writing and speaking task performances,' *Applied Linguistics* 31/2: 236–59.

Yuan, F. and **R. Ellis.** 2003. 'The effects of pre-task planning and on-line planning on fluency, complexity and accuracy in L2 monologic oral production,' *Applied Linguistics* 24/1: 1–27.