Complexity, Accuracy, and Fluency as Properties of Language Performance: The Development of the Multiple Subsystems over Time and in Relation to Each Other

by

Mary Lou Vercellotti

B. A., Carlow University, 1994

M. A., University of Pittsburgh, 2007

Submitted to the Graduate Faculty of

Kenneth P. Dietrich School of Arts and Sciences

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2012

UNIVERSITY OF PITTSBURGH

Kenneth P. Dietrich School of Arts and Sciences

This dissertation was presented

by

Mary Lou Vercellotti

It was defended on

March 16, 2012

and approved by

Dr. Dawn E. McCormick, Lecturer, Linguistics

Dr. Alan Juffs, Associate Professor, Linguistics

Dr. Kevin Hyunkyung Kim, Associate Professor, Psychology in Education

Dr. Nel de Jong, Assistant Professor, Linguistics, VU University Amsterdam

Dissertation Advisor: Dr. Yasuhiro Shirai, Professor, Linguistics

Copyright © by Mary Lou Vercellotti

2012

Complexity, Accuracy, and Fluency as Properties of Language Performance: The Development of the Multiple Subsystems over Time and in Relation to Each Other

Mary Lou Vercellotti, PhD

University of Pittsburgh, 2012

Applied linguists have identified three components of second language (L2) performance: complexity, accuracy, and fluency (CAF) to measure L2 development. Many studies researching CAF found trade-off effects (in which a higher performance in one component corresponds to lower performance in another) during tasks, often in online oral language performance. Trade-off effects are attributed to the inability of the learner to simuletaneously attend to all CAF components at the highest level possible. Although cross-sectional research has suggested that students at different proficiency levels sacrifice performance in one CAF area while improving in another, there has been little longitudinal research about CAF (Ortega & Iberri-Shea, 2005). As such, previous research could not address if CAF grows linearly over time nor if the rate of CAF growth is the same for all learners. The current study explicitly addresses how language performance in CAF changes over L2 development in an instructed environment.

This longitudinal study analyzed English L2 oral data from sixty-six students from Arabic, Chinese, and Korean language backgrounds over 3-9 months in the English Language Institute at the University of Pittsburgh. Elicited speeches were transcribed, coded, and assessed with three measures of structural complexity, a measure of lexical variety, two measures of accuracy, and three measures of fluency. The scores were analyzed with hierarchical linear modeling (Singer & Willett, 2003) to investigate how each student's performance changed over time for each measure and to determine predictive variables. Although individual differences were found in initial scores (often proficiency differences, but not for all measures), growth

trajectories were the same for all measures, except one grammatical complexity measure (length of AS unit) where slopes differed by gender. All measures showed growth, and only two measures (lexical variety and mean length of fluent run) showed non-linear growth. Trade-off effects found in cross-sectional studies were not found in these longitudinal data even though within-individual and between-individual correlations were also calculated. Additionally, the results may suggest that instructed language performance growth is uniform, rather than along individual paths. The research also serves to evaluate the measures, which has research and pedagogical implications.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

This research and much of my graduate research experience was funded by National Science Foundation, Grant Number SBE-0836012 to the Pittsburgh Science of Learning Center (PSLC, http://www.learnlab.org).

I sincerely thank the many people who went on this journey with me and those people who encouraged me to continue, specifically Lauren L. Williford, Claude Mauk, and Natasha Tokowicz.

I must also thank so many people who helped me to accomplish this research: from the PSLC, Anthony Brohan, Michael Bett, David Klahr, Mike Nugent, Ben Madore, Jon-Michel Seman, and Dr. Nel de Jong-my research mentor; from the Psychology in Education Department, Dr. Kevin Kim-my statistics mentor, from the Linguistics Department, Dr. Claude E. Mauk, Dr. Alan Juffs, Dr. Dawn E. McCormick-my ESL pedagogy mentor, and Dr. Yas Shirai-my dissertation mentor, and from my family, Victoria and James, and my husband, Nikolas Seiber, without whom this dissertation research would not have been done.

"What occasion is there then for boasting? It is ruled out. On what principle, that of works? No, rather on the principle of faith."

1.0 INTRODUCTION

Applied linguists have identified three major components of second language (L2) speaking performance: complexity, accuracy, and fluency (CAF). At first blush, complex language is more advanced; accurate language is error-free; and fluent speech is normally paced. However, when looking deeper into each component, these subsystems are complex and multidimensional, and researchers of second language acquisition (SLA) differ on how these components should be defined and operationalized (Housen & Kuiken, 2007). There is support for multiple measures of each component of language performance, and the correlation between measures can substantiate concurrent validity of the measures (Norris & Ortega, 2009).

Since different researchers often use different measurements, comparing results is difficult (Ellis & Barkhuizen, 2005, p. 163). Although testing or measurement research has often been seen as less significant than theoretical research, such research can be beneficial "…to explore the testability of theorectical claims…" (Skehan, 1998a, p. 180). Importantly, accurate measurements allow researchers to consistently observe phenomena, which can be appropriately interpreted and then linked to the theoretical claims about the phenomena (Norris & Ortega, 2003). Further, without standards in the field, reported research may not contribute to the accumulated knowledge because there can be no comparison of the findings (Norris & Ortega, 2003). Chapter 2 reviews the measures for assessing accuracy, complexity and fluency in langauge performance.

Many studies (e.g., Skehan & Foster, 1997; Yuan & Ellis, 2003) researching CAF from instructed language learning settings found trade-off effects in demanding tasks, such as online oral language performance. From a psycholinguistic view of language proficiency, these findings are attributed to the inability of the learner to simultaneously attend to all components of language performance at the highest level possible. As a result, learners must prioritize one component of the language performance. Limited resources are assumed in all three models considered here: limited attentional resources model (Skehan & Foster, 2008), multiple attentional resources model or cognition hypothesis (Robinson & Gilabert, 2007), and dynamic systems theory (de Bot, Lowie, & Verspoor, 2007). However, the field has not reached a conclusion about what components actually trade-off because of differences in tasks, task conditions, and measurements (Ellis & Barkhuizen, 2005, p. 144). Chapter 3 Section 1 reviews the trade-off effects found with between-group designs in the literature, across tasks and task conditions. Importantly, most of the studies that found trade-off effects have looked at learners' performance at a particular time (i.e., performance status) rather than learners' development (i.e., performance change).

Chapter 3 Section 2 reviews the existing literature concerning changes in language performance over time. There has been a call in the field for more longitudinal studies with well-chosen measurements (Norris & Ortega, 2009; Verspoor, Lowie, & Van Dijk, 2008) to better understand development of CAF. For instance, within the construct of grammatical complexity, complexity by subordination is expected to rise and then level off as learners make more use of phrasal complexity (Norris & Ortega, 2009). Although cross-sectional (between-group design) research has pointed to trade-off effect differences based on proficiency, cross-sectional studies can only describe the product not the process of language change (Larsen-Freeman & Long,

1991). There has been little research about growth of CAF components over time (Ortega & Iberri-Shea, 2005). And so, there has been a call for more "micro-developmental studies" (Larsen-Freeman, 2006, p. 614). This study is informed by both information processing theories of language learning and dynamic system theories of studying change in dynamic systems, such as language learning. The current study explicitly addresses the "dearth of research examining what happens as proficiency grows in relation to the performance areas of complexity, accuracy, lexis, and fluency" (Skehan, 2009b, p. 20). Then, Chapter 3 Section 2.3 reviews how individual differences might affect language performance. Chapter 3 Section 3 outlines the research questions that emerge from the literature review.

The methodology is described in Chapter 4, including the participants (Chapter 4 Section 1), data (Chapter 4 Section 2), and the analysis, which focuses on change rather than status (Chapter 4 Section 3). Chapter 5 gives the results of this research by each construct and then across constructs. A summary and general discussion is offered in Chapter 6, and Chapter 7 concludes with the impact of the research and some suggestions for future research, considering the limitations of the current study. Importantly, the field has few true longitudinal studies with a research focus on change over time (Ortega & Iberri-Shea, 2005). As such, our understanding of language development must be inferred based on studies of status. Observing and analyzing actual change in performance can more directly answer theoretical questions about language development. This dissertation fills that gap, focusing on accuracy, complexity, and fluency in language performance over time.

2.0 MEASURING COMPLEXITY, ACCURACY, AND FLUENCY

In each of the components of language performance, there are general measurement and more specific measurements. General measurements can be used in a wider variety of tasks but may not capture differences that a finer-grain analysis could. Specific measurements may capture differences in data related to a particular task in a particular population, but these limit generalizibilty.

Although higher scores on measurements of complexity, accuracy, and fluency may not always be better, (Pallotti, 2009) (e.g., speeches with extremely long utterances without any pauses would not be easily understood), it seems that in general, more is considered better in language performance (Foster, Tonkyn, & Wigglesworth, 2000).

2.1 BASIC UNIT IN SPOKEN DATA

In order to measure frequency of certain forms (such as, clauses) or features (such as errors), researchers must divide the data into consistently defined units. In written data, the text may be safely separated by the student defined units based on the punctuation. Oral language does not have that luxury. Speakers, native and non-native, do not speak in sentences, but in idea units (Luoma, 2004). Oral data usually have many sub-clausal units, especially in unplanned speech (Luoma, 2004). And, this task of dividing data into consistent units is made even more difficult

as speakers hesitate, repeat, abandon topics, and reformulate their speech. For instance, the following is a typical utterance by a low-intermediate adult learner of English. (Pauses are indicated by times in parenthesis. Speech not transcribed as words are preceded by an ampersand. "Ah" indicates a filler of less than 200 milliseconds while "uh" is a filler equal to or greater than 200 milliseconds.) This speech sample is presented again, separated into units after a discussion of the basic units.

[1062] [low-inter] (.75) *&a* (.204) *&a &i ah one day and uh in* (.756) *one day* (.355) *she &ma made it ah* uh *made it for me* (.226) *&i* (1.17) *&s* (.256) *a special uh meal its &ma it uh* (.478) *it's* (.774) *it put in it uh a strawberry and uh* (.723)

Even with a small segment of speech, identifying unit boundaries are difficult. Despite the difficulty in dividing speech into units, determining a basic production unit must be done before determining any frequency measurements. Any measure of length requires that the chosen feature, whether morphemes, words, or characters, be defined by a base production unit.

2.1.1 The (Undefined) Utterance

A major confusion in the literature has been how researchers define "utterance". Unfortunately, some researchers simply label the based unit as an "utterance" or "t-unit" (defined below) and do not define it (Foster et al., 2000). In first language acquistion, dialogic data are often used, and utterance is often equal to "turn". And mean length of utterance, with its roots in 'mean length of response', represents the child's language performance (Parker & Brorson, 2005). However, with adult L2 data, production measures such as mean length of utterance are less reliable, especially since adult learners often use formulaic sequences (Ellis & Barkhuizen, 2005, p. 154). Importantly, in monologic data, an utterance's beginning and end is difficult to identify. If an

utterance is defined by pause bounadaries, complexity with fluency are confounded; if it is defined by syntactic criteria, regardless of intervening pauses, it is a measure of complexity (Norris & Ortega, 2009). Sometimes researchers report the results of mean length of utterance without explaining how the utterances were defined. For instance, David, Myles, Rogers, & Rule (2009) report calculating mean length of utterance, but it is unclear how their utterances are defined. The project description, (available at http://www.flloc.soton.ac.uk/), notes again ambiguously that "Each utterance is transcribed on to a separate line…". By reviewing the actual transcripts, one finds that the lines are defined generally by turn-taking.

When defining "utterance", researchers have employed many units to segment oral data, generally focused on different linguistic levels. Segmenting speech into units based on linguistic criteria is preferred over segments based on a word count because the unit of analysis must have a connection to the psycholinguistic planning process for the measurement to have relevance (Crookes, 1990). Therefore, some researchers use semantic criteria (e.g., proposition, c-unit, idea unit), which seems like a logical way to segment oral data. Defining an "idea unit" and consistently applying that definition to messy data, however, is extremely difficult and is rarely used as the sole criteria for segmenting speech (Foster et al., 2000). Measurements based around idea units are best suited for tasks which have a predetermined, specific content (Ellis & Barkhuizen, 2005, p. 154) so that the text can more easily be compared and coded. Other researchers (e.g., Crookes, 1990) decided on units defined mostly by intonational units (such as tone unit, idea unit with intonation focus, or utterance). Intonation criteria is attractive for oral data, but it is particularly unreliable with non-native speech because learners might not follow the expected intonation patterns of language being learned and because pauses are not reliably an indication of the end of a unit (Foster et al., 2000). Since semantic criteria are difficult to

consistently code and suprasegmental properties are unreliable in learner data, neither is reliable as the main criterion for segmenting oral data.

2.1.2 Defining Utterances by Syntactic Criteria

A third group of units have a mainly syntactic criterion (e.g., sentence, idea unit with a structural definition, t-unit). T-units were found to be most commonly used in the field for both written and spoken data (Foster et al., 2000; Norris & Ortega, 2009), even though it origniated for analysis of the syntactic complexity of written texts. A t-unit was originally defined “one main clause with all the subordinate clauses attached to it” (Hunt, 1965) but was definition was seemingly too vague and was revised to “one main clause plus any subordinate clause or non-clausal structure that is attached to or embedded in it” (Hunt, 1970). Syntactic complexity based on length (words per t-unit) is commonly used in writing research at all levels of proficiency, from elementary school to college writers, with most finding that the mean length of t-unit is correlated with assessments of writing quality (Mills, 1990). When reviewing more recent research using the t-unit, Foster et al. (2000) found that many researchers modified the definition to better fit their data. The main reason the t-unit is modified for oral data is that people do not always speak in full sentences, as expected in writtten text (Luoma, 2004). For example, the c-unit is defined the same as a t-unit, but ellipted answers to questions also count as a clause in oral data (Chaudron, 1988). Modifications of the base-unit, however, make comparisions across studies difficult, and interpretations of any differences uncertain.

2.1.2.1 Analysis of Speech Unit

Foster et al. (2000) suggests a new measure for spoken data, the Analysis of Speech unit (AS unit), which can be reliably and consistently applied to oral data, purportedly without additional modification needed by individual researchers. An AS unit is an utterance with an independent clause (clause with a finite verb) and all subordinate clauses (either clauses with a finite verb or clauses with a non-finite element and at least one other clausal element) associated with the clause.

Foster et al.'s AS unit, following the t-unit, is mainly discerned with syntatic criterion, but intonation and pause information can be used to aid coding. Importantly, clauses with finite verbs separated by pauses greater than 500 milliseconds are generally coded as separate AS units, even if connected by a subordinate conjuction. This adjustment is recommended because some subordinate conjuctions (such as, *because*) function as an ellipted version of an independent clause (such as, *I say this because*…) Although resarchers will still make decisions when attempting to segment and code oral data, Foster et al. (2000) suggest researchers can reliably use their clearly defined unit, and as a consequence, findings can more easily be compared when they are based on the same basic unit.

For instance, the speech example given above is repeated here marked for clauses [^c] and into AS units by lines ended with a period.

[1062] (.75) *&a* (.204) *&a &i ah one day and uh in* (.756) *one day* (.355) *she &ma made*
[low-inter] *it ah uh made it for me* (.226) *&i* (1.17) *&s* (.256) *a special uh meal* [^c].
its &ma it uh (.478) *it's* (.774) *it put in it uh a strawberry and uh* (.7.23)… [^c].

After this review of possible base units and the rational for AS units for this spoken data, the next sections review specific measurements of language performance: accuracy, complexity, and fluency.

2.2 ACCURACY

Accuracy is the most easily defined of the triad since there is more agreement in the goal, which is matching the target language. Housen and Kuiken (2009) define accuracy simply as "error-free" speech. But there is still ambiguity and debate. First, researchers have generally ignored the notion of adequacy in accomplishing a task for the more quantitative accuracy (Pallotti, 2009). Secondly, it is unclear from which dialect the accuracy standards should come. For instance, utterances, such as *I like sport* or *I went to hospital*, could be coded as accurate (in London) or inaccurate (in Pittsburgh), but in practice, a standard must be chosen. Further, assessing accuracy longitudinally may be complicated as learners attempt new lexical items and grammatical forms. As such, Norris and Ortega (2003) caution that accuracy of specific forms may not develop linearly but rather curvilearly. This complication of accuracy of specific forms would make gauaging development more difficult since it might be unclear which part of the arc the data represent. Despite such complications, researchers often measure the accuracy of the language performance of learners.

2.2.1 Measuring Accuracy

Accuracy of the performance can be measured by self-repair attempts or as a function of errors produced (or the lack thereof). Self-repair has been measured as a percentage of self-repairs or as a ratio of self-repairs to errors (Michel, Kuiken, & Vedder, 2007). It is unclear how a lower score or a higher score of self-repairs reflects accuracy in the language produced. Self-correction does not really measure accuracy of the language produced; it is more accurately labeled a measure of the speaker's orientaion toward accuracy (Ellis & Barkhuizen, 2005, pp. 149-150). Given that

accuracy is better measured as a function of errors produced, accuracy can be measured specifically (e.g., accuracy of verb forms) or generally (e.g., overall number of errors or error-free units).

2.2.1.1 Specific Measures

Accuracy is often measured by the learner's suppliance of a specific form in obligatory contexts, which is best suited for focused tasks (Ellis & Barkhuizen, 2005, p. 151). Usually, the researcher decides which form and context to measure based on developmental sequence (proficiency) or task conditions. Often, this means that research with data from lower proficiency students measures accuracy on a different target form than research with data from higher proficiency students because the accuracy on the first target form is expected to approach ceiling. If accuracy is determined by correctness of certain forms, accuracy and development (perhaps complexity or proficiency) are confounded in the measure and potentially misleading (Pallotti, 2009).

When studying language development of a specific structure, learners may be given a task which is likely to elicit the target structure and a coded with a corresponding specific measure of accuracy. Robinson and Gilabert (2007) suggest that specific measures should supplement general measures in order to capture the impact of resource-directing tasks. For instance, when focusing on time and motion, Robinson, Cadierno, and Shirai (2009) used two accuracy measures centered on motion verbs, verb particles, and verb satellites.

Operationalizing accuracy by performance on specific forms, however, does not give a representative picture of the student's overall use of the language. If accuracy is only measured on specific forms, it may not reliability represent the students' general accuracy, especially when analyzing data from students with mixed language backgrounds as certain grammatical features might be easier or more difficult depending on language background (Ellis & Barkhuizen, 2005,

p. 151). Likewise, topic differences may affect specific measures of accuracy in that certain topics may encourage some forms over others. In fact, previous research has found that different topics seem to encourage different forms (De Jong & Vercellotti, 2011). As such, some topics may have few or no instances of the target structure. Thus, the opportunity to use a form and to use the form accurately could differ from topic to topic, which limits the ability to compare accuracy across topics. Further, accuracy scores based on suppliance in obligatory contexts of specific forms may be difficult to interpret (Schachter & Celce-Murcia, 1977), especially student to student from different language backgrounds. Therefore, researchers would have to (very carefully) choose several structures to capture the students' general accuracy.

In addition, measuring accuracy with accuracy on any specific grammatical form ignores any incorrect lexical choices in the language produced (Ellis & Barkhuizen, 2005, p. 139), which means that lexically-based errors would go uncoded in any accuracy measure based on a target structure. Michel, Kuiken, and Vedder (2007) attempted to address these concerns by adding a measure of ommissions (number of article, verb, and subject ommisions per AS unit) and a measure of lexical errors (total number of lexical errors). However, these multiple specific measures can still only capture the errors targeted by the researcher. In conclusion, accuracy based on specific measures of accuracy is usually employed for research on a targeted structure but is less suited to capture overall accuracy performance.

2.2.1.2 General Measures of Accuracy

General measures of accuracy are useful for data from loosely structured tasks where participants have much freedom in responses because students may avoid forms or constructions.

For research other than focused tasks, Ellis and Barkhuzen (2005, p. 151) recommend a general measure of accuracy, such as percentage of error-free clauses or number of errors per 100 words.

Proportion of error-free clauses has been promoted in the field (e.g., Skehan & Foster, 1997). Larsen-Freeman and Long's (1991) recommendation for written learner data is the similar error-free T-units. A variation of this general measure, total errors per AS unit, has also been used (e.g., Michel, Kuiken, and Vedder, 2007). An advantage of using errors per 100 words is that the measure is not complicated by the difficulty of coding a clause, t-unit, or AS unit. However, 100-word segments have no psycholinguistic reality, but segments based on ideas units, clauses, and AS units do (Crookes, 1990).

Further accuracy is most reliably consistently coded at a general level. Schachter and Celce-Murcia (1977) cautioned researchers about the difficulty in classifying an identified error. Often, an ungrammatical sentence can be "corrected" in more than one way because the coder does not definitively know what the speaker intended (Ellis & Barkhuizen, 2005, p. 59). In many cases, the "error" is ambiguous, and the coder makes a subjective decision. For instance, the error (marked with [*]) in the following utterance from an intermediate learner of English could be coded as a missing determiner or as a lack of plural marking on the word "picnic"*:*

> *mmm* (.801) *we (u)sually have picnic* [*] (.204) *in the sea* (.527) *at the sea* (.442) *uh* (.501) *shore* [c] (.3576).

In the next example from another intermediate learner of English, the clause can be coded as a verb agreement error or as a superfluous plural marker on the noun:

> *there's* [*] *no rivers* [c] (1.214).

If a specific measure was used, such as suppliance of determiners or verb form, the data could be skewed by the coding choices.

It is assumed that for the same reason, general measures of accuracy would be less susceptible to first language (L1) influence because as Schachter (1974) showed in her seminal

paper, some constructions may be avoided by some L1 groups. Overall, global measures of accuracy is a more realistic and sensitive measure (Skehan & Foster, 1999).

Moreover, in research that included both general and specific measurements of accuracy, the general measurement has often been found to be sufficiently informative, either giving the same information as the specific measure (e.g., Ahmadian & Tavakoli, 2011; Yuan & Ellis, 2003) or by being more informative than the specific measure (e.g., Michel, Kuiken, & Vedder, 2007). In cross-sectional research comparing the effect of planning, Ahmadian and Tavakoli (2011) found that students in the careful online planning conditions had significantly higher accuracy, measured generally by error-free clauses and specifically by verb forms. Likewise, Yuan and Ellis (2003) found the online planning group produced significantly more accurate narratives than the no planning group, as measured by error-free clauses and by correct verb forms. Michel, Kuiken, and Vedder, (2007), using five accuracy measures, state that only the general measure (number of errors per AS unit) captured differences in language performance between the easier and more difficult information sharing tasks. Although a more specific measure of accuracy did not affect the conclusion, Kuiken and Vedder (2007) report that coding specific error-types gave additional information (i.e., the increase in accuracy was driven by fewer lexical errors). In summary, general measurements of language performance accuracy are better when analyzing data from loosely-structured tasks, for longitudinal data, and when trying to measure general accuracy from students with different langauge backgrounds.

2.3 COMPLEXITY

Complexity has been described as “elaborated language” (Ellis & Barkhuizen, 2005, p. 139). The complexity of produced language has been the most difficult to define and this component of language performance is most easily conflated with language development or progress. By describing complexity as “more advanced” or “challenging language”, it seems as though complexity is not a property of language production but just an indication of development or proficiency (Pallotti, 2009). Complexity can be described relative to proficiency, as “language that is at the upper limit” of the student’s interlanguage system, which is not fully internalized or automatized by the learner (Ellis & Barkhuizen, 2005, p. 139). Skehan and Foster (1997, p. 191) connects complexity with “more challenging and difficult language” or with a “wider repertoire of structures” which is related to “restructuring” of the learners interlanguage. These characterizations have at least three problems. Firstly, with these types of definitions only learners could produce complex language; native speakers with fully internatlized, automatic language would not. Secondly, they seem to conflate complexity with the component of fluency since fluent language is also described as automatic. Thirdly, they seem to wed complexity to recently acquired, but not fully mastered structures. It is also likely that fully mastered structures can be used to produce complex language.

2.3.1 Complexity Measures

After accepting that complexity is a valid component of language performance, researchers have used a myriad of complexity measurements. (See Norris & Ortega, 2009 for a concise review of complexity measurements used in sixteen recent task-based language learning studies.)

Language complexity can be considered a function of sophistication or variety, or a function of syntactic or grammatical complexty (Norris & Ortega, 2009).

2.3.1.1 Complexity by Sophistication

Although complexity by sophistication, based on an acquisition sequence, has been employed in first language acquisition (e.g., Scarborough, 1990), SLA researchers have not frequently attempted to measure this type of complexity (Norris & Ortega, 2009). Pienemann's (1998) Rapid Profile is an SLA version which has been researched as a possible placement test (Spinner, 2011), but it has not often been employed when researching complexity, accuracy, and fluency. One reason may be the number of forms to be tallied is too labor intensive, which is particulary impractical when also looking at the other CAF components. In addition, some complexity by sophistication systems conflate complexity and accuracy when they give partial points to attempted but incorrect structures. However, some researchers have attempted to measure complexity by sophistication simply by choosing a specific construction to tally For instance, Ahmadian and Tavakoli (2011) and Yuan and Ellis (2003) measured syntactic variety based solely on verb forms. Even more specific measurements can be chosen based on the specific research question being tested. For instance, Robinson, Cadierno, and Shirai (2009) measured complexity in part by guaging the complexity of the motion verb clauses.

Neither the full or the simplied complexity sophistication measurements, however, are fully applicable to free-response data. Although these systems may give a description of language performance (i.e., a form was or was not present), looking at specific structures may not yield valid data because of the topic and the freedom of response, which weakens the usefulness of the profile. For example, one topic may encourage discussion of the past and the future, which would increase a measure of syntactic sophisication based solely on verb forms,

but data from other topics may fail to show differences in verb forms. In fact, in research using semi-spontaneous monologues, Recorded Speaking Activity speeches, (which are also used in this dissertation), Spinner (2011) had to exclude elements from the Rapid Profile analysis because the data did not include enough tokens of some forms despite collapsing data across observations. Moreover, Purpura (2004, pp. 36-37) cautions against making assessments following acquistion sequences because the field has not done enough research supporting a fixed order of acquistion. As such, complexity by sophistication is not considered viable for research describing individual growth across observations.

2.3.1.2 Grammatical Complexity

Usually, SLA researchers focus on measuring syntactic or grammatical complexity (Ellis & Barkhuizen, 2005, p. 154). Since adult second language learners can use their advance cognitive resources to create lengthy utterances, Foster et al. (2000) state that it is problematic to measure the grammatical complexity of the speech only by production (length of base unit).

Norris and Ortega (2009) persuasively argue that at least three grammatical compleixty measures (global complexity, phrasal complexity, and complexity by subordination) must be measured since langauge can be elaborated at three different syntactic levels. Global complexity, measured by length of sentence unit (such as words per AS unit), captures complexity in a general sense, in that any addition words, phrases, or clauses will increase this measurement. It seems logical that longer sentences are generally more complex than shorter sentences. In fact, average length is considered to be the best measure of SLA writing (Larsen-Freeman & Long, 1991). A common way for students to make longer sentences is to add a dependent clause. As such, most researchers use a complexity by subordination measure. This type of complexity is most frequently calculated as the mean number of clauses per sentence-length unit (such as AS

unit) or mean number of dependent clauses per total clauses. For instance, syntactic complexity has been operationalized as the ratio of clauses to AS unit in studies (e.g., Ahmadian & Tavakoli, 2011; Michel, Kuiken, & Vedder, 2007) and as the ratio of clauses to t-unit in others (e.g., Yuan & Ellis, 2003; Larsen-Freeman, 2006). On caveat, however, is that complexity by subordination is only valid if the learners have acquired subordination constructions (Ellis & Barkhuizen, 2005, p. 155).

Sentences can also become more complex through modification or nominalization, which can be captured by measuring clause length, or phrasal complexity. Advanced learners are expected to rely less on subordination for increased complexity as they increase phrasal complexity (Norris & Ortega, 2009). Writing research has also found that clause length is a useful measurement in addition to global complexity (Mills, 1990).

Nevertheless, other researchers have used length-based measurements as a measure of fluency (Larsen-Freeman, 2006; Wolfe-Quintero, Inagaki, & Kim, 1998). Wolfe-Quintero et al. departed from previous research by using length-based measurents as fluency measurements. One reason for the change is that as a complexity measure, global complexity does not inform the researcher in how lengthening is achieved. And a second reason is that length-based measurements (such as words per utterance) loaded highly with fluency measures in data (Ortega, 1995, cited in Norris & Ortega, 2009). Norris and Ortega (2009), however, strongly criticize the use of length based measurments as measures as fluency. They were unconvinced by Wolfe-Quintero et al.'s arguments for the radical departure from previous research, and rightly so. Obviously, using the two additional measurements of language complexity does give insight on how legthening is achieved. In response to the factor-analysis reasoning, Norris and Ortega explain that the length-based measurements cited were utterance-based defined by pause

information, which should not be used as a complexity measure. The authors specifically caution against using utterance as a denominator in any complexity measure because then complexity is conflated with fluency. Therefore, length-based measurements which are based on syntactic criteria, such as AS-units, are clearly measurements of complexity. To sum, three measures of syntactic complexity have been endorsed, global (words per AS unit), subordination (clauses per AS unit) and subphrasal (words per clause). These three measures of grammatical complexity, in conjuction, are expected to capture the construct of complexity across proficiency levels. Emperical studies investigating this claim is reviewed in Chapter 3 Section 2.1.

2.3.1.3 Lexical Variety Measures

Lexical choices also constitute a facet of language complexity. Lexical variety can be calculated by finding the type-token ratio (TTR), which is the number of word types divided by all word tokens. Variation of this type of measure abound. For instance, some researchers count the number of different word families or the ratio of functional words to lexical words (Ellis & Barkhuizen, 2005, p. 155). Since TTR is highly influenced by the size of the corpus, lexical variety is better measured by a more complex procedure (Malvern & Richards, 1997). One such measure is the Guiraud adjustment of TTR, which adjusts the TTR for the text size by substituting the number of tokens in the equation with the square root of the tokens (types/√tokens). Another adjustment for text length effects, the mean segmental TTR (MSTTR), determines the mean TTR of (50-word or ten-word) segments of the text.

A more complicated adjustment is D. Since TTR falls at a predictable rate as the text size increase, D scores compare the lexical variety found in the text to the theoretical models (McKee, Malvern, & Richards, 2000). Importantly, it can reliably compare texts of different lengths because the software runs multiple trials on groups of randomly selected words. In a

study of French L2 learners and French bilinguals, Treffers-Daller (2009) found Guiraud's and D scores both adequately distinguish between groups, although the authors conclude that D slightly more powerful measure for their data. The measurements of lexical complexity described so far are text-internal measurements as they each use only the text itself to determine lexical complexity.

Researchers might also want to measure the lexical complexity considering the relative frequency of the words, which has been called lexical sophistication (Read, 2000). One alternative measure of lexical complexity, P_Lex, suggested by Meara and Bell (2001), is based on the number of "hard" words in each ten-word segment. However, there are several assumptions about frequency and lexical difficulty that must be addressed. Using external norms for any language preformance measure is problematic since it is unclear what the standard should be (Ellis & Barkhuizen, 2005, p. 56). First, it is unclear which corpora should determine the frequency. Meara and Bell used Nation's (1984) word list for researching written texts. Skehan (2009a) used the spoken language section of the British National Corpus to calculate P_Lex scores. Second, after choosing an external measure, it is unclear what frequency threshold separates "easy" from "hard" words. Meara and Bell considered all words in Nation's first 1000 words "easy" and any words beyond the 1000 most frequent as "hard". Skehan chose 150 per million words as the threshold between easy and "difficult" words. In fact, both Meara and Bell and Skehan admit that their chosen thesholds are arbitray. Since neither researcher reported a testing of other thresholds, it is an open question. It is also debatable if frequency is a valid determinant of "difficult" lexical items. As Meara and Bell point out, frequency does not always indicate difficulty or even appropriateness. Additionally, P_Lex scores, or any text external measure are more affected by topic effects than text internal measures, in that a specific topic

may encourage a “hard” word which is very frequent in that particular topic (i.e., a certain narrative might encourage a particular less frequent word). For instance, a narrative about a visit to a fortune-teller might have a high P_Lex score if “fortune-teller” is used often, even if the text does not have much internal variety. If most students use the same word throughout the task, it seems unsound to label it “difficult”. As such, a text external measure of lexical variety is less useful across topics.

2.3.2 Summary of Complexity Measures

Although many measures have been used to capture language complexity, SLA researchers have converged on grammatical complexity with multiple subparts and lexical variety. Grammatical complexity measures should capture the multiple ways learners can increase their language performance: a measure of global complexity to gauge the average length of the base unit (e.g., AS unit for spoken language), a measure of phrasal complexity to capture increases within the clause (such as modification), and a measure of subordination to capture increases by adding dependent clauses. Lexical variety has also been proposed as a crucial part of language complexity. There is more support for internal measures of lexical variety (based on the types and tokens in the sample itself) than external measures (considering a relative frequency or difficulty of words used). Lexical variety measured with D, a sophisticated TTR measure which adjusts for the length of text, has been increasingly used in the field.

2.4 FLUENCY

Fluency is commonly used in a broad sense, similar to second language proficiency, such as "She's fluent in French" (Koponen & Riggenbach, 2000). However, in this paper, fluency has a more narrow meaning, as a component of language performance, specifically the "delivery of speech" (Schmidt, 1992, p. 358). The component of fluency is especially prone to be assessed holistically by raters. However, the raters can be biased, of course, influenced by the student's accuracy as well as temporal-based fluency measures (Schmidt, 1992). And, raters can be especially vulnerable to responding to their individual construct of fluency (Koponen & Riggenbach, 2000), such as the broad sense of fluency, which may include lexical choices, grammatical complexity, and pragmatics. Therefore, researchers have attempted to more specifically define and measure fluency separately than holistic ratings.

2.4.1 Fluency Measures

When analyzed into subcomponents, fluency has been discussed in terms of repair, speed, breakdown of fluency, and automatization (Skehan, 2009b). As discussed in Section 2.1, repair has also been described as a measure of accuracy. Generally, self-corrections were not considered predictive of fluency (Schmidt, 1992) and are not included here. Since self-corrections are confounded with accuracy and length of utterances are confounded with complexity, fluency measures should capture temporal variables of oral language performance.

Speed of language performance is probably best captured by speech rate, which is calculated as the total number of syllables divided by the total time, or articulation rate, which is calculated as the total number of syllables divided by the total time, excluding filled and unfilled

pauses. Articulation rate in L2 speech has been shown to increase over time in an intensive language program (De Jong & Perfetti, 2011) and in study-abroad programs (DeKeyser, 2007). However, these measures are not germane in this study for several reasons. First, a speaker can have a high speaking rate or articulation rate but may not be fluent, since neither measures the number or length of pauses. Second, speaking rate and articulation rate are highly subject to individual difference as evidenced by high correlations between L1 and L2 speaking rates (Towell, Hawkins, & Bazergui, 1996). Furthermore, talking quickly is not the point, as speaking teachers will stress. A non-fluent speaker is not specifically identified by slow speech but by a breakdown of fluency. For instance, the fluency section of the analytic scale assesses fluency by descriptors such as "smooth flow" and "pausing…is very evident", not speech rate (Council of Europe, 2001). Although articulation rate can capture the amount of lengthened syllables, which may be similar to filled pauses, this is of minimal concern in this study. Finally, speech rate captures the amount of language similarly to phonation time ratio, which is discussed below.

2.4.1.1 Fluency Breakdown - Pausing

A breakdown of fluency is indicated by pauses, either the number of pauses, length of pauses, or the placement of pauses. In fact, the University of Pittsburgh's English Language Institute uses a grading rubric for the Recording Speaking Activity (RSA) speeches which attends to both the amount of pausing ("few", "some", "many") and seemingly to the placement of the pauses (no, some, or many "unnatural pauses") (English Language Institute at the University of Pittsburgh, 2007). Some researchers suggest that the location of pauses is relevant because second-language learners are more likely to pause mid-clause whereas native speakers pause at clause boundaries (e.g., Skehan & Foster, 2008; Crookes, 1989), seemingly because online planning cannot wait until a clause boundary. This observation has led some researchers (e.g., Skehan, 2009b) to

propose a ratio of clause boundary pauses to mid-clause pausing. Yet, a student can have an otherwise "fluent" speech even if the pauses are not syntactic boundaries (Towell, Hawkins, & Bazergui, 1996). And specifically, within phrase pausing is particularly disfluent and non-native, but that has not been incoporated into a pause-location model, presumably because it is too cumbersome to measure. As such, temporal measures of fluency are more common. In fact, temporal measures are considered the basis of fluency (Ellis & Barkhuizen, 2005, p. 140).

Rather than placement of pauses, a more global measure of fluency is warranted. Phonation time ratio gives a general or global view of a speaker's fluency as it is simply the percentage of time the student spoke during the recording. There are three main ways that a speaker can increase her phonation time ratio in a speech, by decreasing the number of pauses, decreasing the length of pauses, or by increasing the length of speech between the pauses. The number and length of the pauses both reflect the extent the speaker must pause to plan their language performance (Ellis & Barkhuizen, 2005, p. 156). In native speech, pausing is attributed to "attentional preoccupation with micro-planning" (Schmidt, 1992, p. 377). In addition to the general measure of phonation time ratio, which captures both pausing and speech production information, more specific measures of fluency breakdown, such as mean length of pause (MLP) adds insight into the fluency of the performance. The average length of pause is expected to decrease as fluency increases, obviously.

2.4.1.2 Fluency Proceduralization - Mean Length of Fluent Run

The speech between pauses is labeled mean length of run or mean length of fluent run (MLFR). With native speakers, these stretches of speech are assumed to reflect "skilled micro-planning that does not require much attention" (Schmidt, 1992, p. 377). Previous research has found that mean length of run best captured differences in fluency during a narrative task (Towell,

Hawkins, & Bazergui, 1996). This increase in length of a uninterrupted speech is thought to be a result of the proceduraliztion of knowledge, as described in Anderson's Adaptive Control of Thought (ACT) model (1983).

Either a lower mean length of pause or a higher mean length of run would indicate higher fluency. If a speaker simply decreases the number of pauses, while maintaining the length of pauses and the length of speech between pauses, phonation time ratio itself will capture the higher fluency (De Jong & Perfetti, 2011). Therefore, a combination of these three measures (phonation time ratio, mean length of pause, and mean length of fluent run) can adequately capture differences in fluency in oral performance.

2.5 SUMMARY OF MEASURING CAF

In this chapter, I reviewed how the constructs of complexity, accuracy, and fluency in language performance have been measured. The AS unit is a useful base unit for spoken data because it offers a consistent coding scheme for spoken data which have restarts, reformulations, and oral language fragments. Although specific and general measures have been used in language research, general measures seem to be more practical for coding longitudinal data with different topics from a heterogeneous population. Grammatical complexity and lexical variety have both been considered relevant subcomponents of language complexity (Skehan, 2009b). Since utterances can be expanded by adding clauses or by adding modifiers within clauses, multiple measures of grammatical complexity is warranted, in addition to a general measure of grammatical complexity. As such, Norris and Ortega (2009) recommend at least three non-redundant measures of grammatical complexity: one of general length (such as, length of AS

unit), one of subclausal length (such as, length of clause), and one of subordination (such as clauses per AS unit). Oral fluency can be improved by decreasing the number of pauses or the length of pauses, or by increasing the length of utterances between pauses. Therefore, at least three (non-redundant) measures of fluency should capture improving fluency, such as phonation time ratio, mean length of pause, and mean length of fluent run, respectively. Overall, accuracy, grammatical complexity, lexical variety, and fluency measures are expected to capture language development in the speech of adult second-language learners.

3.0 RELATIONSHIPS BETWEEN COMPLEXITY, ACCURACY, AND FLUENCY

After choosing measurements for complexity, accuracy, and fluency, research have often considered if and how these constructs of language performance interact. Section 3.1, offers a review of studies that looked at potential trade-off effects (where a higher performance in one component corresponds to a lower performance in another) between complexity, accuracy, and fluency during langauge performance. The review focuses on research of oral language performance, and particularly in learners of English.

Researchers often design cross-sectional studies to conduct quantitative analysis of trade-off effects across proficiency levels, since data can be collected from many participants in a short period of time. Such research studies language performance at a point in time (i.e., the learner's language performance status). Norris and Ortega (2009) suggest researching how these three constructs interact with longitudial studies. A longitudinal design allows the process of language learning to be followed over time (Larsen-Freeman & Long, 1991). Aggregating data across particpants of differing proficiencies is often assumed to adequately substitute for longitudial results, but that assumption is debateable (Larsen-Freeman & Long). The research reviewed in Section 3.2, such as Larsen-Freeman (2006) are longitudinal designs, but these papers have often studied written language performance. Comparisons, therefore, are tenuous since oral language performance has different attentional demands than written language performance, and written data, obviously, can not capture a measurement of oral fluency.

3.1 TRADE-OFF EFFECTS IN LANGUAGE PERFORMANCE

Many researchers accept limitations in learner language performance. Simply put, focusing on one component of language performance might result in a lower performance in one or both of the other components. From a cognitive, information processing framework, Skehan (2009) predicts a competitive relationship between CAF because of limited mental resources, specifically limited attentional capacity and working memory. In Skehan's limited attentional model, this limited capacity during online processing is a result of a single-source view of attention. If trade-off effects are expected, a theory should predict which CAF constructs are likely to show the effects and why. Skehan (1998a, p. 286) states that adult learners emphasize meaning over form, which can potentially hinder further language development. Then,when learners do focus on form, there is a secondary contrast between control of form (accuracy) and interlanguage risk-taking (complexity). All language learners have these tensions during performance. When a performance shows improvement, rather than trade-off effects, in two areas, Skehan (2009b) suggests there are two possible explanations. The growth in two areas could actually the result of separate influences. For instance, the task structure may aid accuracy while the information manipulation during the task requires the students to use subordination which increases grammatical complexity. Alternatively, when analyzing group data, some individuals may attend to one area of the CAF triad while others attend to another area. Aggregated data may then give the appearance that two areas, which should be in a competitive relationship, are both showing improvement. Therefore, correlations must be run on individual performances, not just at group levels, he suggests.

Even for researchers who reject a single-source capacity limitation, trade-off effects may be found in language performance, but these trade-off effects can be explained by attention

control and interference (Robinson, 2003). Robinson's cognition hypothesis (Robinson & Gilabert, 2007, p. 162) claims that increased accuracy and complexity can be encouraged by increasing the cognitive demands of the task given to learners. As the students attempt to produce the language required by the greater functional demands in the relatively increased complexity of the task, their language performance will improve. The cognition hypothesis has pedagogical implications for designing and sequencing tasks, from simple to complex. The related framework categorizes "task complexity" (based on cognitive factors), "task conditions" (interactive factors), and "task difficulty" (based on learner factors). The task design can either direct resources (this does not hinder performance) or disperse resources (which does hinder performance). Specifically, learners can produce greater accuracy and language complexity during complex tasks that are resource-directing (e.g., talking about more elements rather than a few elements).

From a dynamic systems theory approach, 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 dynamic systems theory or the similar complexity theory (Larsen-Freeman, 2012) approach reject a cause-and-effect model of language learning (de Bot, Lowie, & Verspoor, 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, Lowie, & Verspoor, 2007). In addition, competitive relationships are likely to be temporary (Van Geert, 2008). When resources are interlinked, limited resources do not always result in trade-off effects (de Bot, Lowie, & Verspoor). Some subsystems might show supportive growth (Larsen-Freeman, 2009). In a "supportive" relationship, growth would be found in both areas of performance. Specifically, less

advanced to more advanced measures are more likely to be in supportive relationships (Van Geert, 2008). Supportive, rather than competitive, relationships are theoretically possible despite limited resources because "connected growers" require fewer attentional resources than unconnected subsystems. Therefore, a key to this theoretical approach is which subsystems have meaningful relationships (Verspoor, Lowie, & Van Dijk, 2008) and what relationships are more advanced.

In summary, it is assumed that learners cannot generally attend to all aspects of language performance because the online processing demands are greater than learners' capacity. Therefore, learners prioritize their language performance, resulting in trade-off effects based on the task and their orientation (Ellis & Barkhuizen, 2005, p. 140). In contrast, dynamic systems theory does not assume trade-off effects will necessarily result. Tasks given to learners are also expected to affect language performance; specifically, a more complex task will push learners to accomplish more. The next section reviews the existing research concerning expected trade-off effects from the task.

3.1.1 Trade-off Effects Predicted in Language Performance

Regardless of the theoretical framework, some trade-off effects (Table 1) are expected. Specifically, Skehan's trade-off hypothesis expects tension between meaning (usually measured as fluency) and form (either complexity or accuracy). Skehan's understanding of adult language learners motivates this trade-off. Skehan (1998a, p. 269) states that adult learners vary in learning style by learning through exemplars and emphasizing fluency or by learning through analysis and emphasizing complexity or accuracy. This meaning versus form dichotomy has also been studied as a limitation in attending to information, such as VanPatten's (2007) input

processing theory. Pedagogy also echos the fluency-form distinction, in which spontaneous, free-flowing language is the goal of fluency-oriented tasks and a focus on form and control is the goal of accuracy-oriented tasks (Brumfit, 1984). Empirical findings have supported the form-meaning dichotomy. In a study looking at the effect of task repetition, Bygate (2001) found that grammatical complexity increased but at the expense of fluency (measured by the number of pauses).

As mentioned earlier, accuracy and complexity may compete during oral language performance. Skehan and Foster (1997) reported a trade-off between accuracy and complexity in a study looking at the effect of planning during three oral tasks. The planning group had higher means on all measures than the non-planning group. During the decision-making tasks, the planning group significantly outperformed the non-planning group on the complexity measure but not the accuracy measure while on the narrative task, the planning group significantly outperformed the non-planning group in accuracy but not in complexity. Importantly, during the personal information task, the planning group significantly outperformed the non-planning group on all three measures.

Robinson's (2001b) cognition hypothesis, however, does not predict a trade-off in complexity and accuracy. It predicts that in resource-directing tasks, a more complex task will result in an increase in the complexity and accuracy in the language performance of that task. Specifically, in simple monologic tasks, fluency is likely to be promoted (but not complexity or accuracy), while accuracy and complexity (but not fluency) are promoted during complex monologic tasks (Robinson, 2001a, p. 307). When testing the cognition hypothesis, Michel, Kuiken, and Vedder (2007) found that students performing the more difficult task had increased accuracy but a decrease in fluency (driven by the dialogue condition) with no significant effect

on language complexity. Yuan and Ellis (2003) also reported an accuracy and fluency trade-off within the careful online planning condition. Michel, Kuiken, and Vedder conclude that the task "seems to direct the learner's attention" (p. 254). Skehan (2009) recognizes that certain tasks seem to alleviate some tension on attentional resources, such as personal information tasks generally have higher accuracy and fluency, and pre-task planning allows learners to produce language with more complexity and fluency.

Table 1 Empirical Findings Showing Trade-off or "Competitive" Effects

researcher(s)	design	task	participants	trade-off effects
Ahmadian & Tavakoli (2011)	between groups; 1-way ANOVA	oral narrative about video	intermediate English L2, Persian L1, adult females (n = 60)	accuracy (error-free clauses; verb forms) vs. fluency(# of syllables/min. of speech =PTR; pruned PTR) with COLP
				complexity (subordination; syntactic variety) vs. fluency with COLP
Yuan & Ellis (2003)	between groups; 1-way ANOVA	oral narrative about cartoon	English L2, Chinese L1 undergraduates (n =42)	accuracy vs. fluency with COLP
				accuracy vs. lexical complexity with OLP
Michel, Kuiken & Vedder (2007)	2 X 2 (+/- few elements, +/-mono)	oral info. sharing task	intermediate Dutch L2 from Turkey and Morocco (L1s not given) (n = 44)	accuracy vs. fluency (only in combined monologic and dialogic conditions)
Skehan & Foster (1997)	2 X 2 (planned/ unplannedpost-task/ no post-task	oral task (personal information, narrative, decision-making task)	pre-intermediate English L2, mixed L1 adult (n = 40)	accuracy (proportion of error-free clauses vs. complexity (clauses/c-units)
Skehan (2009a)	between group	various oral tasks	low-intermediate English L2	lexical complexity (D) vs. grammatical complexity -subordination
				lexical complexity (P_Lex) vs. accuracy (error-free clauses)
				lexical complexity (P_Lex) vs. grammatical complexity -subordination

A recent study (Ahmadian & Tavakoli, 2011) found higher accuracy and grammatical complexity (contrary to Skehan's prediction of a form tension) at the expense of fluency. This finding of improvement in two constructs of language performance was attributed to the task design. Students who are encouraged to do careful online planning (COLP) when describing a

cartoon-based narrative had higher accuracy and grammatical complexity than students in the pressured online planning condition, who had higher fluency. Thus, this between-group research found strong performances in both accuracy and grammatical complexity at the expense of fluency, which does support a meaning form tension but not a secondary tension within form. In Robinson's cognition hypothesis, the contrast between these groups might be understood along the resource-dispersing dimension, where the online planning group has to split their attention between accuracy and complexity, which would lower both.

Lexical complexity further complicates the form-accuracy trade-off hypothesis. Yuan and Ellis (2003) reported a lexical variety (MSTTR) and accuracy trade-off in the oral production of narratives. However, when looking at text-external lexical variety measures, lexical variety (P_Lex) was reported by Skehan to be negatively correlated with accuracy and somewhat negatively correlated with grammatical complexity. In general, the relationships between lexical complexity and other components of language performance are still unclear.

3.1.2 Connected Growers

As discussed above some studies (Ahmadian & Tavakoli, 2011, Skehan, 2009, Skehan & Foster, 1997) which found trade-off effects, also showed correlated components and are entered in Table 2. Another research study, Mizera (2006), showed that accuracy and fluency seemed to be connected growers since the speed fluency measure (number of syllables) was negatively correlated with number of errors while number of errors and number of pauses measure was positively correlated. In each pair, producing fewer errors was correlated with improved fluency. This study's finding, found by comparing scores within-individuals, is contrary to multiple findings in cross-sectional research designs.

Lexical complexity has been reported to show both a competitive relationship (in Section 3.1.1) and supportive relationship with global grammatical complexity, and with accuracy. In a review of his own research, Skehan (2009a) reports that for non-native speakers, lexical variety (measured by D) is positively correlated with accuracy (measured in error-free clauses) and negatively correlated with grammatical complexity (measured by subordination). David, Myles, Rogers, and Rule (2009) found lexical variety (Guiraud's Index) significantly correlated with global grammatical complexity when aggregated across age groups. Note that the findings concerning lexical complexity are mixed.

Table 2 Correlated Components "Connected Growers" in Language Performance

researcher(s)	design	task	participants	correlated components
Ahmadian & Tavakoli (2011)	between group	oral narrative of video	intermediate English L2, Persian L1, adult females, (n = 60)	accuracy & grammatical complexity-subordination with OLP
Mizera (2006)	correlations of language status	oral narrative of picture book	Spanish EFL; English L1 (n = 44)	accuracy (number of errors) and fluency (number of syllables)
David, et al. (2009)	cross-sectional (3 groups x 2 years apart)	information sharing conversations	adolescent, French L2; English L1 (n = 60)	lexical complexity (Guiraud TTR) & global complexity (mean length of utterance in morphemes)
Skehan (2009)	between group	various oral tasks	low-intermediate English L2	lexical complexity (D) & accuracy
Skehan & Foster (1997)	2 X 2 (planned/ unplanned; post-task/ no post-task	oral task (personal information, narrative, decision-making task)	pre-intermediate English L2, mixed L1 adult (n = 40)	accuracy (proportion of error-free clauses &. complexity (clauses/c-units) & fluency with planning in a personal information task

3.1.3 Context of the Previous Findings

The context (task instructions or task and research design) in which the language performance was given should be considered when reviewing conclusions about language performance or development.

3.1.3.1 Task Instructions

The direct impact of the task instructions given to the participants must inform conclusions about trade-off effects. In some studies which report trade-off effects, the cross-sectional design might have induced a difference in focus during language performance. For instance, Yuan and Ellis (2003) studied the effect of planning on oral language performance and concluded that there was a trade-off effect between accuracy and fluency based on group score comparisons. The trade-off effect, however, was not found *within* each planning group. The group (online planning) with the lowest fluency did have the highest accuracy. The group (pre-planning) with the highest fluency, however, did not have lowest accuracy; and the group (no planning) with the lowest accuracy did not have the highest fluency. Yuan and Ellis (2003) conclude the lack of a consistent trade-off effect is because the planning groups used their planning time differently, in that pre-task planning is used for fluency but online planning is used for accuracy. It is unclear that their conclusion is valid considering that the online group was specifically encouraged to attend to accuracy with the following instruction: “If you think you say something not correct or not to your satisfaction, you can correct it as many times as you can.” As such, the fact that the online planning group had higher accuracy but lower fluency is not surprising. It is unclear why the pre-task planning group, if using the planning solely for fluency, did not suffer in accuracy; the pre-

task planning group had statistically similar accuracy scores as the "accuracy focused" online planning group. Importantly, within-individual correlations, which could illuminate if individual students did sacrifice performance in one construct over another, were not reported.

In fact, in a similar study design, Ahmadian and Tavakoli (2011) begin to question the trade-off effect assumption, noting that "the simultaneous use of careful online planning and task repetition positively impacts the EFL learners' accuracy, complexity and fluency" (p. 56). Their conclusion is that the online planning helped the complexity and accuracy while the repetition allowed the fluency to improve. Rather than question the trade-off hypothesis, they attribute the lack of expected trade-off effects to separate influences of the task design. As Skehan (2009b) points out trade-off effects may be obscured by separate influences of task design. Therefore, research with more complex data analysis is needed to uncover information obscured in group means.

3.1.3.2 Effect of Task and Research Design

The existing research has found a variety of competitive trade-off effects (Table 1) and connected growers (Table 2) from varying tasks and task conditions from different experimental designs. However, it must be noted that many of competitive effects were inferences from a cross-sectional design and based on group-score comparisons, which may not represent the performances of the individuals. For instance, Skehan and Foster (1997) conclude there were trade-off effects because the difference between the planning group and the non-planning group did not reach significance for complexity (p = .10) for the decision-marking task nor for accuracy (p = .14) in the narrative task. Moreover, they state that their findings offer "very strong" evidence of trade-off effects, even though the planning group outperformed the non-planning group on every task, and the planning group significantly outperformed the non-planning group

in complexity (p = .01), accuracy (p = .02) and fluency (p = .001) during the information-sharing task. As such, the trade-off was at the study population level (between two different tasks), not even at the group level (i.e., *within* each task). Further, they reported only group means across tasks and planning conditions, rather than looking at how individuals performed, so it is not clear that trade-off effects would be found at the individual level. Similarly, Ahmadian (2011) states his repeated measures research with students in two conditions (massed repetition and control) supports Skehan's trade-off hypothesis simply because his repetition group increased their complexity scores, particularly words/AS unit, but not their accuracy scores, including error-free clauses. Inferring a trade-off effect is unfounded for at least two reasons. First, both groups in Ahmadian's study had very similar accuracy rates during the pre- and post-tests. Therefore, the improvement in complexity did not come at the expense of accuracy (i.e., accuracy did not decrease with the increase in complexity). Second, the pre-test task and intervention task was a narrative, but the post-test was an interview task, which may have required different, perhaps novel, grammatical structures. As such, it is plausible (if not more likely) that the accuracy scores in both the pre- and post-test simply represent a baseline for a novel task for that population. Caution is warranted with making conclusions about possible trade-off effects with results from cross-sectional research designs with group score comparisons, rather than looking for trade-off effects within individual performances during tasks.

3.1.4 Summary of Trade-off Effects

Although the findings differ and sometimes contradict, some trade-off effects in language performance are anticipated. Particularly, a trade-off between accuracy and fluency seems to be a robust finding (Ahmadian & Tavakoli, 2011; Yuan & Ellis, 2003; Michel, Kuiken, & Vedder,

2007) although there is one study (Mizera, 2006) with the opposite finding. In a study comparing the effect of planning across tasks, complexity and accuracy seemed to be in a competitive relationship during two of the tasks (Skehan & Foster, 1997). Also, research with between-group designs (Ahmadian & Tavakoli, 2011; Yuan & Ellis, 2003) has found students can have higher accuracy and complexity at the expense of fluency. An emerging key explanatory variable is the task or task instruction given to the student. The demands of the task or instructions can encourage the learner to prioritize one component of the triad over the others (Ellis & Barkhuizen, 2005, p. 143). Previous literature supports that tasks or task instructions can induce performances with "trade-off effects" based on the focus of the task, but these findings may not reflect the limitation of performances because of limited attentional capacity. It is unclear if the students must prioritize or how students will prioritize the CAF components of language performance, without the effects of the differing demands of the task or task condition. A second key question involves whether trade-off effects, common in the literature from cross-sectional research, will be found when looking at individual performances. A single study (Mizera, 2006), which measured the language performance status within-individuals, did not find trade-off effects but a growth in both accuracy and fluency. Little work has been done to research language performance development (change) rather than performance status. In addition, it is unclear how proficiency affects the trade-off effects in language performance. The next section reviews research that looks at how CAF language performance may change over time as proficiency increases.

3.2 LANGUAGE PERFORMANCE OVER TIME

Since most research has been done at a single time-point (which measures the status of the participants' performance) with a homogenous proficiency group, it is unclear how CAF constructs of language performance change over time. It may be that conflicting results in the literature may be at least partially explained by different relationships at different proficiency levels. Cross-sectional research infers the effect of proficiency with group selection, but other differences may exist between groups which are nevertheless attributed to the difference in school level or proficiency. Therefore, studies with repeated measures can better investigate how CAF in language performance changes over time. Section 3.2.1 reviews the research addressing growth within a single construct. To date, only the growth of grammatical complexity is predicted by theory, and most studies have reported findings about growth within that construct (cross sectional and repeated-measures designs). Section 3.2.2 reviews the longitudinal research concerning growth of CAF across constructs. Section 3.2.3 reviews research findings about individual differences in language performance growth.

3.2.1 Growth within CAF Constructs over Time

Researchers from a dynamic systems theory framework generally have theories regarding relationships *within* a CAF construct, rather than across CAF constructs, which is more the focus in research from an information processing framework described in Section 3.1.1. In a dynamic systems theory framework, less advanced stages support more advanced stages, whereas relationships from more advanced stages to less advanced are competitive (Van Geert, 2008). For example, for children learning an L1, single-word utterances support the growth of multi-

word utterances; but as multi-word utterances increase, single-word utterances will decrease. This dynamic would be found in all levels of development and in all areas. Therefore, this theory would predict a supportive relationship from the less complex stages to the more complex, but a competitive relationship from the more complex to the less complex.

Consistent with a dynamic systems theory framework, Halliday and Matthiessen (1999) state that learners express ideas initally by individual words, clauses, and sentences, then these are expanded by subordination, and then by grammatical metaphor. So, in the measurement of grammatical complexity, as phrasal complexity increases, complexity by subordination will decrease. So far, empirical evidence, as discussed below, supports the grammatical complexity sequence.

In a written English L2 cross-sectional study, global complexity (length of t-unit) linearly increased at each higher level of proficiency while clause/t-unit ratio increased but then decreased at the highest level of proficiency (Flahive & Snow, 1980). Likewise, another large-scale cross-sectional study of texts written by Chinese L1 learners of English found that global complexity (mean length of sentence) and phrasal complexity (mean length of clause, which was defined by a subject and a finite verb) significantly increased linearly with proficiency while complexity by subordination (clauses per sentence) significantly decreased linearly with proficiency (Lu, 2011).

In addition to cross-sectional research, repeated measures research has also supported this developmental sequence within grammatical complexity. First-year Japanese college students significantly increased the complexity by subordination (clauses/sentence) in English L2 writing after one semester, whereas the increase in global grammatical complexity (words/sentence) was not significant (Wendel, 2007). Conversely, EFL students in a massed repetition condition

increased scores on global complexity (words/AS unit) whereas their subordination scores (clauses/AS units) did not significantly increase after six months (Ahmadian, 2011).

In a case study of a single Finnish learner of English, written homework assignments indicate a competitive relationship between a specific measure of phrasal complexity (based on length of noun phrase) and sentence complexity (based on averaging the number of dependent clauses) (Spoelman & Verspoor, 2010). Some further support for the developmental sequence of grammatical complexity was found in a case study of a single Dutch learner of English. In this study, two specific measures of grammatical complexity (noun phrase length and finite verb ratio) were found to be connected growers (Verspoor, Lowie, & Van Dijk, 2008). Specifically interesting is that the learner lengthened sentences by increasing phrasal complexity (in the noun phrase) in the later assignments. Table 3 summarizes the empirical findings related to growth among grammatical complexity measures.

Some of these findings support the dynamic systems theory concept of simultaneous development or connected growers. Complexity at the word level is a connected grower with phrasal complexity (noun phrase) and with sentence complexity (based on mean number of dependent clauses (Spoelman & Verspoor, 2010). Global complexity (mean length of sentence) and phrasal complexity (mean length of clause) were significantly correlated (r = .571) in the English L2 essays written by Chinese university students (Lu, 2011), although this may be simply because an increase in phrasal complexity can increase global complexity. Unfortunately, these studies, could not test if the developmental sequence of grammatical complexity in language performance is significantly influenced by language background. Further, although dynamic systems theory predicts the relationships within a developmental sequence of this single construct, the theory has not yet offered a developmental sequence within the other constructs or

across constructs. As was reviewed in Section 3.1, trade-off effects have been found in between-group studies, but it is unclear if or how these might change over time which will be reviewed in the next section.

Table 3 Empirical Findings about the Growth of Grammatical Complexity

researcher(s)	design	task	participants	finding
Flahive & Snow (1980)	cross-sectional based on proficiency	English L2 texts	six levels of proficiency (L1 not reported) (n = 300)	global grammatical complexity increases at each level
				complexity by subordination increases at levels 2-5, but decreases at level 6
Lu (2011)	cross-sectional based on university level	timed written English argumentative essays	Chinese L1, English L2; 4 levels (n= 412)	global complexity (mean length of sentence) increases linearly
				phrasal complexity (mean length of clause) increases linearly
				complexity by subordination (clauses/sentence) decreases linearly
Wendel (2007)	repeated measures (2 observations 8 months apart)	written English narratives based on cartoon	Japanese L1, English L2 (n = 36) ; first-year university students	global complexity (words/sentence) did not significantly increase
				complexity by subordination (clauses/sentence) significantly increased
Ahmadian (2011)	cross-section and repeated measures (6 months)	Oral narratives and oral interview	Iranian L1 English L2, (n=15 in each group) ntermediate, females	global complexity (words/AS unit) increased pre- to post-test, with massed repetition group outperforming control
				Complexity by subordination (clauses/AS unit) increased insignificantly in both groups
Spoelman & Verspoor (2010)	repeated measures (over 3 years)	English L2 homework assignments	English L2, Finnish L1 (n=1)	phrasal complexity (noun phrase) competitive with sentence complexity
Verspoor, Lowie & Van Dijk (2008)	repeated measures (over 3 years)	written academic English	advanced English L2, Dutch L1, (n=1)	complexity (number of words/finite verb ratio correlated with specific phrasal (NP)
				phrasal complexity increased in latest text

3.2.2 Growth across CAF Constructs over Time

Research about the development of CAF in language performance is scarce since research from information processing frameworks has not generally employed longitudinal designs to investigate complexity, accuracy and fluency. Researchers working in dynamic systems theory or complexity theory have begun to collect empirical data in longitudinal designs looking for relationships between the components of language performance, but, unfortunately, longitudinal research tends to use written texts rather than spoken data, as was seen in the previous section. The few longitudinal research studies reviewing CAF performance includes two dynamic systems theory studies with written data (i.e., Spoelman & Verspoor, 2001; Verspoor, Lowie & Van Dijk, 2008) and a single study with oral and written language CAF performance (Larsen-Freeman, 2006).

In written language performance, global complexity (length of sentence) seems to have a slight competitive relationship with lexical complexity (TTR), especially in the middle time points (Verspoor, Lowie, & Van Dijk, 2008). Another study with written homework assignments found no meaningful relationship between accuracy and complexity (Spoelman & Verspoor, 2010).

In her longitudinal study of five learners of English in the People's Republic of China, Larsen-Freeman (2006) found that each construct (global complexity, complexity by subordination, accuracy, and lexical variety) showed growth, albeit statistically insignificant, based on group averages after six months, but each student's pattern of growth differed. Larsen-Freeman reported two "preferred paths": a focus on lexical variety or a focus on grammatical complexity, based on visual inspection of the data. This qualatiative analaysis, however, was not

supported with correlations, specifically, within-individual correlations from the muliple measures which could indicate if student was choosing to focus on one area over another.

When Higgs and Clifford (1982) attempted to explain why certain students with a specific speaking performance profile did not show language performance growth, they found "terminal" or "fossilized" students have relatively high fluency and vocabulary scores, but low accuracy scores. Higgs and Clifford suggest that this group of students fail to progress beyond a 2/2+ score (out of five) because of proactive interference, in which learning how to communicate infers with the ability to subsequently learn how to communicate with accuracy. Likewise, Ahmadian (2011) concludes that the participants in his research attended to complexity (length of AS unit) at the expense of accuracy, based on an increase of AS unit length but no increase in percentage of error-free clauses. It must be noted, however, that the control group also showed no increase in accuracy. In summary, Larsen-Freeman's research on written English L2 suggests that all CAF constructs grow although not necessarily by the same route, while Higgs and Clifford's discussion on spoken language growth (English L1) and Ahmadian's English L2 research suggest that particular routes of development may stop further development across CAF constructs.

Table 4 summarizes the findings concerning growth across CAF. Obviously, more longitudinal research is needed in order to evaluate if all CAF constructs continue to progress or if specific patterns of growth limit progress across CAF. In fact, Higgs and Clifford (1982, p. 76) specifically state that "the existence of the terminal profiles must be independently verified". It is unclear if some (or all) students have similar language performance at the beginning and at the end of the language program.

Table 4 Growth across CAF Constructs

researcher(s)	design	task	participants	finding
Verspoor, Lowie & Van Dijk (2008)	repeated measures (over 3 years)	written academic English	advanced English L2, Dutch L1, (n=1)	global grammatical complexity competes with lexical complexity, especially at middle proficiency
Spoelman & Verspoor (2010)	repeated measures (over 3 years)	English L2 homework assignments	English L2, Finnish L1 (n=1)	accuracy not related to complexity
Larsen-Freeman (2006)	repeated measures (4 observations over 6 months)	written English personal narrative	Chinese L1, English L2, female, 27-27 years (n = 5)	lexical variety (types/√(2) tokens) vs. grammatical complexity-subordination (clauses/t-unit)
Ahmadian (2011)	Repeated measures (6 months apart)	dialogic English narrative and interview	Intermediate EFL, Iranian L1, English L2, female, 18-21 , (n=15)	Global complexity (words/AS unit)increases but accuracy does not
Higgs & Clifford (1982)	experiential data	oral language performance	English L1, foreign language learning	students with high vocabulary and fluency did not show growth in grammar

3.3 INDIVIDUAL DIFFERENCES IN LANGUAGE PERFORMANCE AND GROWTH

Section 3.1.1 showed that the task, the task condition, and the design of the research often induced participants to prioritize one language component over another. Such research with between-group designs, however, accepts that the learners are basically the same and any differences are caused by the treatment, but differences in the rate of change or the route of change might be systematically related to other causes. Research has found language learning differences based on individual differences, e.g., L1, gender, age, motivation (Romaine, 2003).

3.3.1 Affective factors

Individual differences based on affect (e.g., language aptitude, self-consciousness, assertiveness) can influence language performance when the task and planning condition are held constant. Social factors, such as motivation (Dörnyei & Skehan, 2003) and assertiveness (Ockey, 2011), are considered influential to language development, particularly in oral performance. For instance, Ockey (2011) found that assertiveness, but not self-consciousness was found to be an explanatory variable in oral performance. Measurements of participants' language aptitude, self-confidence, or other affective measures are not available for this study are not the focus here. It is noteworthy, however, that extraversion has been connected to increased fluency in L2 performance but not necessarily to increased accuracy (Dewaele & Furnham, 1999).

3.3.2 Age

Age at the time of testing, the information available here, has not been found to be predictive of scores; most studies contribute age differences to age of onset. (Hyltenstam & Abrahamsson, 2003). Age of onset can either be considered by start of studying a language or immersion in an English speaking environment. Age of onset, measured by time studying a language, may not meaningful here because of the possibility of differing ideas of what "studying" the language means to each student. Time spent in an English speaking environment might be more predictive for some aspects of language performance. Most studies looking at the length of immersion included people with over ten years of immersion. Age of immersion in the L2 environment has often been cited as a contributing factor about level of attainment, but overall age has not found

to change the route of language learning, (Ellis R. , 1994, p. 491). Overall, age is not a focus in this study of adult second-language learners.

3.3.3 Gender

In a review paper, Wallentin (2009) concludes that sex is not a significant predictor in language proficiency. Cameron (2009) also argues that gender differences in language ability are not accepted by most language researchers (But see Baron-Cohen, Knickmeyer, & Belmonte, 2005 for a discussion of sex differences in the brain which may result in language deficits). Males and female students, however, might have different rates of growth because of unequal learning opportunities (Romaine, 2003).

3.3.4 Initial Proficiency

It is expected that higher proficiency students will have better initial scores. It is unclear, however, if initial proficiency upon enrollment in an intensive English program affects growth rate. Larsen-Freeman (2006) found different rates of change among her five homogenous learners of English. In reading ability, learners of higher proficiency have been found to improve more quickly, (Stanovich, 1986). Alternatively, lower proficiency learners might have faster rates of growth if they have lower starting point upon enrollment and have similar outcomes. Wendel's (2007) findings of foreign language learning support greater growth for students with lower initial proficiency. However, in a repeated measures study, Kuiken and Vedder (2007) found no interaction between proficiency and measure of performance, suggesting that proficiency will not affect rate of change.

3.3.5 Language Background

Most SLA researchers accept cross-linguistic influence (positive and negative transfer) affects L2 learning (Odlin, 2003). For example, while learning English articles, Spanish L1 and Italian L1 students have been found to use no article before acquiring target-like usage whereas Chinese L1 have been found to use a demonstrative pronoun before acquiring target-like usage (Zobl, 1984). Luk and Shirai (2009) show that L1 affects the timing of English morphemes, specifically that Japanese, Korean, and Chinese learners of English tend to acquire the possessive marker ‘*s* earlier than Spanish learners of English but acquire the plural marker -*s* and articles later than Spanish learners.

Research with general measures of accuracy, however, have not found L1 differences in language performance. Advanced learners of English from five different language backgrounds (Arabic, Chinese, Korean, Malay, and Spanish) did not significantly differ in grammatical complexity by subordination (clauses per T-unit) or in global accuracy (errors per clause) in written texts (Bardovi-Harlig & Bofman, 1989). This finding is hardly conclusive, however, since it included only six learners from each language. Bardovi-Harlig and Bofman did find individual differences, and the group means differed between L1 group and between proficiency groups, though not statistically significantly. (Interestingly, the lower proficiency students had a higher mean complexity score, which the authors did not address.)

It should also be noted that it is difficult to separate the effect of language background from cultural background. Students from a shared language background share more than just an L1. Learning environments may create culturally-based learning styles or at least contextual learning styles (Wong, 2004).

3.3.6 Learner Orientation

Other individual differences may also influence the language performance. For instance, individual differences on a linguistic focus may exist, that is whether the learner prioritizes complexity, accuracy, or fluency during language performance. Larsen-Freeman (2006) suggests that the learner's interlanguage path is influenced by the learner's L1, the L2, and the learner's orientation (to complexity, accuracy, or fluency). Variation in the rate of growth has been found in individuals in L1 acquisition (Bates, Dale, & Thal, 1995), and variation in the route of development has also been suggested (Wells, 1986). Therefore, differences might be expected to continue in individuals during L2 development, in both rate of growth and development path.

Skehan (1998a, pp. 269-270) suggests that the interaction between learning opportunities and language learning aptitude creates three paths of language growth. Learning style and the choices associated with the style have advantages and disadvantages (Dörnyei & Skehan, 2003). The ideal learner would balance the goals of fluency (meaning), complexity (interlanguage risk-taking), and accuracy (language control) so that all three constructs would develop as connected growers (Skehan, 1998a, p. 269). As such, the ideal learner would aim to acquire new forms, would gain control over the forms, and would integrate the form so that it can be performed fluently. Meisel, Clahsen and Pienemann (1981) included a related notion of motivation in their multidimensional model and suggested that learners with an "integrative orientation" value reaching the target language norms. In motivation research, this behavior might indicate a mastery-approach (Elliot & Murayama, 2008) as it prioritizes learning the language. Balanced learners attend to both form and meaning, and this balanced approach may be the key to language learning (R. Ellis, 1994, p. 549).

Conversely, unbalanced learners are either overly analytic or overly communicative. Analytic learners may achieve high complexity but have difficulty in fluent language performance. When learners prioritize accuracy, they value control of their language performance, whereas learners prioritizing complexity are willing to attempt challenging language or a variety of different structures (Skehan, 1998b). This orientation might be reflected in a language performance with higher grammatical and lexical complexity perhaps at the expense of accuracy since they are willing to attempt difficult constructions. Over time, these students are expected to have higher complexity and accuracy. According to Skehan (1998a, p. 270), however, these analytic path learners would need pedagogic pressure to gain fluency.

In contrast, communicative learners have acquired fluency earlier in the process which hinders progress in other areas of language performance. Learners who value fluency may focus on meaning over form (Skehan, 1998b). These learners may focus on meaning by simplifying the complexity of the language performance, such as Meisel, Clahsen and Pienemann's (1981) "segregative orientation". Ellis and Barkhuizen (2005, p. 139) suggest that learners who prioritize meaning (fluency) will avoid or solve linguistic problems quickly. Alternatively, it might be expected that learners who prioritize fluency, might have lower accuracy. According to Skehan (1998a, p. 270), these learners risk fossilization, as Higgs and Clifford (1982) suggested for students with higher vocabulary and fluency skills relative to grammatical skills.

These alternative paths in performance are the result from the choices the language learners' make since learners cannot attend to all facets of CAF during online language production. Larsen-Freeman's research also pointed to "distinctive orientations and paths" (2006, p. 601). Limited memory and other attentional resources induce trade-off effects.

Some students balance these trade-off effects and are expected to show more even growth. Other learners are unbalanced, either by sacrificing fluency while improving complexity or accuracy, or by sacrificing accuracy (and complexity) while improving fluency. Some previous research has found differences in learning behavior, which impacted language performance (Politzer & McGroarty, 1985). Politzer and McGroarty found Hispanic students and Asian students reported significantly different learner behavior during an intensive English course. These cultural differences in language learner behavior may reflect differences in previous English instruction (Skehan, 1998a, p. 269) or cultural characteristics (Wong, 2004).

It is unclear if proposed paths of interlanguage development will be evident from the individual differences in CAF performance over time. Additionally, it is unclear if age, initial proficiency, gender, L1, or instruction cohort (the demographic variables considered in this study) or any other individual differences systematically affect rate of growth across CAF.

3.4 SUMMARY OF LANGUAGE PERFORMANCE

Language performance over time can investigate growth patterns within constructs, growth patterns across constructs, and individual differences in language performance.

Researchers within a dynamic system theory have suggested that the interconnected nature of language development means that relationships are likely to be found within the sequence of development (de Bot, Lowie, & Verspoor, 2007; Van Geert, 2008). Grammatical complexity has been shown to grow first though adding dependent clauses, which would be found by a measure of complexity by subordination, and then through more complex language, which would be found by a measure language complexity at the phrasal level. Little work,

however, has been done thus far to support such relationships, save for the development of grammatical complexity in written texts (Wendel, 2007; Spoelman & Verspoor, 2010; Verspoor, Lowie, & Van Dijk, 2008).

The interconnectedness of language development may also reach across constructs. In other words, the development (i.e., change) of one construct can affect the development of another. For instance, Larsen-Freeman (2006) suggested that focusing on improving lexical variety may mean ignoring grammatical complexity. Higgs and Clifford propose that language learners who are sufficiently proficient to communicate (lexical proficiency and fluency) do not continue to develop grammatical accuracy. Since Higgs and Clifford's paper relied on antidotal evidence and Larsen-Freeman's paper relied on in-depth look at five students, these proposals need to be tested with larger quantitative studies.

Individual differences have been found for language performance (status) and language performance over time (change). Language background has been associated with differences in performance at a given time while affective factors, age, and gender have been suggested as possibly affecting the rate of change. Researchers have suggested that initial proficiency and learner orientation may affect both status and change rate of language performance.

3.5 SUMMARY OF THE ISSUES

There are two main issues: the possibility of trade-off effects and language performance over time. Each issue is summarized in the following sections before a summary of the remaining issues which culminates into the articulation of the research questions.

3.5.1 Trade-off Effects

It is accepted in the field that learners cannot attend to all areas of CAF performance, especially in demanding tasks. Although the findings differ and sometimes contradict, some trade-off effects in language performance are anticipated. Many studies with cross-sectional research designs report a trade-off between accuracy and fluency, while Mizera's (2006) findings, based on correlations of individuals' performance, suggest these are connected growers. Also, research with between-group designs (Ahmadian & Tavakoli, 2011; Yuan & Ellis, 2003) has found students can have higher accuracy and complexity at the expense of fluency. However, an emerging key explanatory variable is the task given to the student and the research design.

3.5.2 Language Development

Language performance over time can investigate growth patterns within constructs, growth patterns across constructs, and individual differences in language performance. Since language development is complex and likely interconnected, relationships within constructs (e.g., a developmental sequence of grammatical complexity) and among constructs are likely to be found, but research is needed to support that theoretical proposal. More research is needed to support the existence of preferred paths to development and even terminal paths found with experiential impressions.

Individual differences have been found for language performance (status) and language performance over time (change). Language background has been associated with differences in performance at a given time, but most longitudinal studies have generally had homogenous participants, such as female, high intermediate English learners from the People's Republic of

China (Larsen-Freeman, 2006) or first-year university students in Japan learning English (Wendel, 2007). Researchers have suggested that initial proficiency and learner orientation may affect both status and change rate of language performance.

3.5.3 Remaining Issues

Little work has been done to research language performance change rather than language performance status. More longitudinal research, especially with oral language, is needed to observe change in language performance.

It is unclear if CAF performance develops fairly similarly or if there are multiple paths of development of CAF in instructed language learning settings. Moreover, since the longitudinal studies (Larsen-Freeman, 2006; Spoelman & Verspoor, 2010; Verspoor, Lowie, & Van Dijk, 2008; Wendel, 2007) have generally had homogenous participants, it is unclear if variation can be explained by the available demographic information or if there are systematic patterns in growth across CAF.

Some research studies, as reviewed in Section 3.2, have used longitudinal designs with written language data (Larsen-Freeman, 2006; Spoelman & Verspoor, 2010; Verspoor, Lowie, & Van Dijk, 2008; Wendel, 2007) in order to explore relationships within and across constructs, but it is unclear if research on oral language performance will follow results of written learner data. Moreover, while relationships between constructs in oral language performance have been found in cross-sectional research, longitudinal research is needed to confirm (or dispute) that relationships on a task performance (status) continue over time. In other words, it is unclear if the relationships among the subsystems of language performance (CAF) impact language development.

3.5.3.1 Research Questions

With a focus on longitudinal oral performance data from students with mixed language backgrounds, my research questions are as follows:

RQ1: Is there significant individual growth in learners' performance over time? I hypothesize that all measures will increase over time, following previous research with written data (Larsen-Freeman, 2006; Wendel, 2007) but contrary to Higgs and Clifford (1982) description of a "terminal" spoken language profile.

RQ2: Do individual differences explain individual growth? Obviously, participants with higher initial proficiency are expected to have higher initial scores. I hypothesize that students with lower proficiency will see steeper rate of growth since they have more room for improvement, following Wendel (2007). I do not expect other demographic factors in the study (i.e. age, gender, L1) to be predictive on these general measures of complexity, accuracy, and fluency.

RQ3: What are the relationships between the CAF language performance measures? It is unclear if patterns of learner growth described by Higgs and Clifford (1982) or Skehan (1998a) will be found in the data. As stated above, I hypothesize that all measures will show improvement, but trade-off effects are expected even though most of the trade-off effects findings were based on studies of status rather than change because all of the considered theories assumed limited resources. Specific hypotheses follow.

RQ3a: **What is the relationship between the three measurements of grammatical complexity?** I hypothesize that complexity by subordination will rise and then plateau and will have a negative correlation with phrasal complexity in the latest speeches, which

supports the dynamic systems theory, as predicted by Halliday and Matthiessen (1999) and was found by in previous research on English L2 writing (Flahive & Snow, 1980).

RQ3b: What is the relationship between grammatical complexity and fluency? I predict a positive correlation between the three measures of grammatical complexity and fluency because participants may be able to much produce language fluently (without regard to accuracy).

RQ3c: What is the relationship between grammatical complexity and accuracy? I hypothesize some students will show a relatively strong negative correlation between grammatical complexity and accuracy, as learners will make errors as they attempt structures at the higher edge of their proficiency. This finding would support Skehan's (1998a, p. 286) expectation of tension between control and risk-taking in language learning.

RQ3d: What is the relationship between fluency and accuracy? I hypothesize a significant negative correlation between the measures of fluency and accuracy since the participants have little chance for pre-task planning and must rely on pressured online planning, even though the existing research found this trade-off only in careful online planning conditions (Ahmadian & Tavakoli, 2011; Yuan & Ellis, 2003). This finding would support Skehan's (1998a, p. 286) expected tension between meaning (fluency) and form (accuracy).

RQ3e: What is the relationship between lexical variety and grammatical complexity? Following Larsen-Freeman (2006), I hypothesize a negative correlation between complexity by subordination and lexical variety. I predict a positive correlation between lexical variety and phrasal complexity since phrasal complexity is expected to be higher at higher proficiency levels. I hypothesis that lexical variety and global grammatical complexity will be negative at lower proficiency levels since the use of varied lexical items may reduce resources to

attend to grammar at lower proficiency levels. This prediction is tentative considering the conflicting existing results about the relationship between lexical variety and global grammatical complexity found in the literature (Skehan, 2009a).

RQ3f: What is the relationship between lexical variety and fluency? I hypothesize a negative correlation between measures of fluency and lexical variety, especially at the lower levels, since the retrieval and articulation of lexical items may cause fluency breakdown.

RQ3g: What is the relationship between lexical variety and accuracy? I predict a positive correlation between lexical variety and accuracy, indicating as lexical variety increases, accuracy increases since both signal a greater control (rather than production) of the language.

In summary, I hypothesize all measures, generally, will show growth over time in the intensive English program. Initial proficiency is expected to affect initial scores (i.e., students with higher initial proficiency will have higher initial scores) and affect growth of CAF (i.e., students with lower proficiency will have greater gains). Some CAF measures are expected to be positively correlated (lexical variety and phrasal grammatical complexity, lexical variety and accuracy, and grammatical complexity and fluency). Some CAF measures are expected to be negatively correlated because of developmental sequence (complexity by subordination and phrasal complexity) while some are expected to be negatively correlated because of the learners' limitations in attentional resources (grammatical complexity and accuracy, fluency and accuracy, complexity by subordination and lexical variety, global complexity and lexical variety, lexical variety and fluency). The next section outlines the methodology to analyze the development of complexity, accuracy, and fluency and the relationships between the measures.

4.0 THE STUDY

4.1 METHODOLOGY

This research was a descriptive-quantitative design, as are many longitudinal SLA studies (Ortega & Iberri-Shea, 2005). This study included the coding and analysis of two-minute semi-spontaneous monologues from the Recorded Speaking Activity (RSA). These recorded speeches are often part of the curriculum of the speaking classes in the English Language Institute at the University of Pittsburgh. The RSAs were part of the speaking class curriculum in every semester of 2010 calendar year but were not part of the curriculum during the spring semester of 2011.

4.2 PARTICIPANTS

This study's population was limited by the population in the English Language Institute at the University of Pittsburgh during 2010. In order to expand the scope of the research to be as generalizable as possible, data from two instruction cohorts (described below) were studied. The English Language Institute program and placement procedure were determined to be sufficiently standardized to warrant pooling data from successive instruction cohorts. Although the inclusion of two cohorts increased the variability in the data (e.g., different topics, different semesters, etc.), the benefits of the increased statistical power with additional participants outweighed the

disadvantages from the increase in variation. Data from participants with at least three speeches or from two semesters were included. The mean number of observations per participant was 4.45 (SD = 1.3), and 37.9% of participants had exactly four speeches (Table 5). 62.1% of the participants gave either four or five speeches. In general, the average enrollment time is two semesters (range one to four semesters). Therefore, the data fairly represent the span of a typical student's enrollment.

Table 5 Frequency of Number of Observations per Participant

Number of Observations	Frequency Percentage
3	22.7%
4	37.9%
5	24.2%
6	1.5%
7	13.6%

This research was limited to students from three largest language backgrounds, Arabic (Gulf Arabic and Libyan Arabic[1]), Mandarin Chinese (Taiwan and People's Republic of China), and Korean. There were sixty-six participants in two recent proficiency level cohorts who met the minimum requirement of at least three speeches or two semesters of enrollment. The demographic information about each participant was limited to the information the program collects from the participants. Specific and reliable[2] information about the beginning of language studies is not available because the demographics questionnaire asked for only broad information about length of language learning, for example, "less than 1 year", "1-2 years", "3-5 years, and "more than 5 years". Nonetheless, differences in length of studying English should be captured

[1] The three Libyan Arabic speakers were males in cohort 2. In a post-hoc review of the data, these participants were not outliers in demographic information (independent variables), and did not pattern as a subgroup in any of the language performance scores (dependent variables).

[2] I found inconsistent information reported by some students, such as, reporting "more than 5 years of English learning and "more than 5 years" in an English environment but reporting "less than 1 year" and "1-2 years" the following semester.

by the score of initial proficiency during the placement testing described in Section 4.3.1 below. Considering time spent in an English speaking environment, only four of the 66 participants (6%) reported more than five years in an English environment, with an additional four students reporting three-five years. This indicates that most students (nearly 88%) reported less than three years in an English speaking environment. Importantly, these data also indicate that all the participants entered an English environment as adult second language learners.

The demographic information of each instruction cohort is described in Sections 2.1 and 2.2 respectively, followed by a comparison of the two instruction cohorts a comparison by language group, and a summary of the participants overall.

4.2.1 Instruction Cohort 1

In this study, cohort 1 was enrolled at the high-intermediate level for speaking class during the summer 2010 semester. Some of these students were enrolled in the low-intermediate speaking class during the spring 2010 semester; some continued to the advanced level during fall 2010. Cohort 1 had twenty-four students, fourteen male and ten females; seventeen with an Arabic language background, six with a Chinese language background, and one with a Korean language background. The mean placement test score (to be described in Section 4.3.1), the listening test, was 18.8 (SD = 4.95), with a range of 9 - 26. The means per language group (Arabic M= 18.7, SD = 4.9; Chinese M= 20.2, SD = 4.6); Korean M = 21, SD n/a) were not significantly different, $F(12, 23) = 1.440$, $p = .277$. Additionally, 41.7% of the students in cohort 1 tested into the low-

intermediate level and 58.3% tested into high-intermediate level[3]. The mean age of cohort 1 was 24.8 years (SD 4.4) with a range of 19 – 33 years.

4.2.2 Instruction Cohort 2

In this study, cohort 2 was enrolled at the high-intermediate level for speaking class in fall 2010. Some of these students were enrolled in the low-intermediate speaking class during summer 2010. Although some may have continued to the advanced level speaking class in the spring 2011, the RSA speeches were not part of the speaking curriculum that semester. Cohort 2 was larger with forty-two students, twenty male and twenty-two female, but demographically similar with twenty-six with an Arabic language background, ten with a Chinese language background, and six with a Korean language background. The mean placement score as a measure of initial proficiency was 19.5 (SD = 4.31) with a range of 11 - 27. Again, the language groups within this cohort (Arabic M = 17.6, SD = 4.8; Chinese M= 22.0, SD = 2.8; Korean M = 21.5, SD = 1.4) were similar, $F(14, 41) = 1.948$, $p = .067$. Similar to cohort 1, 40.5% of cohort 2 tested into the low-intermediate level while 59.5% tested into the high-intermediate level. The participants' mean age in this cohort was 25.7 years (SD = 4.6) with a range of 18 – 35.

4.2.3 Comparison and Summary of Cohorts

Overall, the participants (Table 6) had a mean age of 25.3 years (SD = 4.5), and a mean initial proficiency score of 19.2 (SD = 4.5). The mean age and proficiency scores were similar between

[3] Although students may test into the advanced level, such students were not included in this study because they were not in the program long enough to provide three or more observations.

the two cohorts, which was confirmed by a two-tailed t-test for age, $t(64) = -.647$, $p = .520$ and initial proficiency scores, $t(64) = -.828$, $p = .411$. The participants were almost split evenly by gender: males (n = 34), females (n = 32). See Appendix C for specific individual demographic information. Participants vary in number of observations (speeches) resulting in 166 observations (56.5%) from male students and 128 observations (43.5%) from females.

Table 6 Summary of the Participants' Demographic Information

demographic	Summary of participants		
age	mean 25.3 years (SD. = 4.5), range 18-35 years		
gender	males (n = 34)		
	females (n = 32)		
initial proficiency	mean placement test score = 19.2 (SD = 4.5), range 9-27		
L1	Arabic (n = 43)	Gulf Arabic	Saudi Arabia (n = 35) Kuwait (n = 5)
		Libyan Arabic	Libya (n = 3)
	Chinese (n = 16)	Mandarin	China (n = 5)
			Taiwan (n = 11)
	Korean (n= 7)		South Korea (n = 7)

4.2.4 Comparison of Language Background Groups

The participants were not evenly distributed by language background (Table 7). There were more students from an Arabic language background (n = 43) than all other language backgrounds (n = 23). There were 208 observations (70.7%) from Arabic students and 86 observations (29.3%) from non-Arabic students because there were more Arabic-speaking participants (65%). In addition, there were more observations per Arabic student (M = 4.84, SD = 1.23) on average than Chinese (M = 3.75, SD 1.13) or Korean student (M = 3.71, SD = .49).

The three language groups, however, were similar in initial proficiency and age. The language groups had similar group means on the placement test (Arabic M = 18.05, SD = 4.8; Chinese M = 21.38, SD = 3.6; Korean M = 21.43, SD = 1.3), $F(17, 65) = 1.541$, $p = .120$. The

language groups also had similar mean ages (Arabic M = 24.42, SD = 4.2; Chinese M = 26.92, SD = 4.0; Korean M = 27.43, SD = 5.8), $F(2, 65) = 2.848$, $p = .065$. The Chinese and Korean language groups were more homogenous than the larger Arabic group in terms of intial proficiency.

Table 7 Participant Information by Language Background Group

	mean	standard deviation	range
Number of Observations			
Arabic (n=43)	4.84	1.2	3-7
Chinese (n=16)	3.75	1.1	3-7
Korean (n=7)	3.71	.49	3-4
Initial Proficiency			
Arabic (n=43)	18.05	4.8	9-27
Chinese (n=16)	21.38	3.6	13-25
Korean (n=7)	21.43	1.3	20-23
Age			
Arabic (n=43)	24.42	4.2	18-34
Chinese (n=16)	26.92	4.0	19-34
Korean (n=7)	27.43	5.8	19-35

4.2.5 Language Background

The participants have three different language backgrounds, Arabic, Chinese, and Korean. This section serves to give a general overview of the basic grammar of each language and reviews research on English learning issues of each group, including the context of learning English. The focus of this study is development of the CAF constructs generally. Acquisition of specific structures or error analysis is not the focus here, and specific error types are not being analyzed for the present study. As such, first language (L1) interference cannot be postulated as the cause of any errors, so only general overviews of the language grammars are offered. This failure to address specific errors and possible L1 transfer or linguistic influence does not imply the denial

of L1 interference, especially given that some L1 influence, especially in the form of learned attention seems inevitable (N. C. Ellis, 2006). Rather, for this study there is no a priori reason to expect that one language group would be more or less affected by L1 interference on this task. Regardless, language background was limited to the three largest L1 groups in order to consider L1 as possible predictor.

It is important to note that although avoidance of structures might result in higher accuracy in the short-term, Ortega (2009, p. 40) suggests that low risk-taking students using an avoidance strategy might delay language development because those errors were avoided.

4.2.5.1 Arabic Language Background English Learners

Arabic is a Semitic language of the Afro-Asiatic language family (Aoun, Benmamoun, & Choueiri, 2010, p. 1). Eloquence, which includes repetition, is a highly valued trait in Arabic culture (Nydell, 2006, p. 97). Arabic speakers learn a regional dialect of Arabic (e.g., Gulf, Maghreb, Egypt, Levant) as a native language and learn Modern Standard Arabic, which is used for writing and formal speaking, in school (Aoun, Benmamoun, & Choueiri, 2010, p. 2). Mahmound (2000) concludes that language transfer comes from both dialects (although the dialects share many features), based on error analysis in translation and free writing task.

Arabic is subject-prominent language. The basic word order is verb (V), subject (S), object (O), but other word orders (VSO, and VOS) are possible (Aoun, Benmamoun, & Choueiri, 2010) since Arabic has case markings. Arabic also has extensive subject-verb agreement, including person, number, and gender (Aoun, Benmamoun, & Choueiri, 2010). A copula is not needed in present tense clauses. Arabic has modifiers following the noun (e.g., relative clauses) without a subordination marker (e.g., relative clause) but an obligatory pronominal reflex (e.g., resumptive pronoun) (Schachter, 1974). In Arabic, the absence of a

definite article conveys indefiniteness so indefinite articles are usually unspoken and unwritten (Thompson-Panos & Thomas-Ruzic, 1983).

Researchers have found that Arabic students learning English have difficulties with verbs, prepositions, articles, and relative clauses (Scott & Tucker, 1974). Thompson-Panos and Thomas-Ruzic (1983) contribute Arabic students' difficulty with articles and verbs (e.g., lack of copula) to L1 transfer. Schachter (1974) reported that Arabic students produce many English relative clauses (similar to native speaker controls) but have a relatively high percentage (20%) of errors in the structure. Since Arabic does not have the equivalent of a relative pronoun for relative clause constructions, Arabic speakers may use both the relative clause with a subject pronoun or the clause may be a separate clause (Thompson-Panos & Thomas-Ruzic, 1983). In this study, if Arabic students employ the first strategy, it would result in an error-marked clause. The second strategy would result in an additional AS unit.

English is taught at all level in the Gulf region, usually by native Arabic speakers before the university level (Syed, 2003). Syed reports that the explosive growth of education system, particularly English as a second language, in the Gulf region has negatively impacted the quality of the education.

4.2.5.2 Chinese Language Background English Learners

The family of Chinese languages is from the Sino-Tibetan branch of languages (Chen, 1999). Mandarin is the standard dialect of both the People's Republic of China and Taiwan, and although regional dialects exist, the differences are mainly in phonology (Lin, 2001, p. 1). Mandarin is a tonal language (Lin, 2001, p. 44).

Mandarin is a topic-prominent language. The basic word order is SVO with few morphological affixes (Lin, 2001). Compounds are common. Mandarin does not have case markers or a copula. Noun modifiers precede the noun. There is no article system in Chinese.

Schachter (1974) concluded that Chinese students avoid relative clauses by making paraphrase clauses. If Chinese students employ that strategy in the study described here, each clause would be counted as separate AS units in this study.

Memorization is a favored learning strategy in China, for learning Chinese characters and for learning English (Chen, Warden, & Chang, 2005). Chinese learners of English have few chances to talk with native English speakers, and the highly important exams do not measure oral communication (Chen, Warden, & Chang, 2005).

4.2.5.3 Korean Language Background English Learners

Korean has been considered related to Japanese (Sohn, 1999, p. 11) but is also considered a language isolate. Standard Korean is the central dialect, around Seoul, and there are few grammatical differences among the geographic dialects (Sohn, 1999, p. 12). Korean has intonational stress groups.

Korean is a topic-prominent language. Its basic word order is SOV, with many morphological markers (Yang, 1994). The copula is found only with predicate nouns (Lee, 1989, p. 43). Korean has no article system. Verbs include the stem and suffixes to mark inflection (e.g., tense, aspect, modality as well as other grammatical information, such as, nominalization, and adverbials (Yang, 1994). Korean, however, does not have subject-verb agreement (Cho, 2004). Korean relative clause constructions come before the noun and are not introduced with a relative clause but are marked with an adjectival affix while resumptive pronouns are optional (Yang, 1994).

Cho (2004) comments that Korean learners of English struggle with subject-verb agreement, prepositions, and articles required in English. Cho also suggests that Koreans avoid the postmodified relative clauses in English by separating complex sentences into two simple sentences.

English education in secondary school emphasizes reading and grammar, in preparation for entrance exams (Cho, 2004). Neither the students nor the public school teachers are fluent speakers of English, but English proficiency is valued (Cho, 2004). Korean students tend to be afraid of making mistakes and "may respond in short phrases because they may not feel confident" (Cho, 2004, p. 35).

4.3 THE ENGLISH LANGUAGE INSTITUTE

The English Language Institute at the University of Pittsburgh is an intensive English program for adults. Full-time students are usually simultaneously enrolled in speaking, listening, grammar, reading, and writing classes, but enrollment in all five classes is not required. Some students may have classes at two instruction levels, such as high-intermediate speaking, but low-intermediate grammar. This research focused on the speaking course. The program usually has multiple classes of each subject area at three instruction levels (the low-intermediate, high-intermediate, and advanced). Generally, the students are enrolled in the intensive English program in order to pass the TOFEL exam so that they can begin undergraduate or graduate classes at an English speaking university. Some students enroll to improve their English skills for their profession or for everyday living in an English speaking country.

4.3.1 Placement Tests

Upon acceptance into the English Language Institute program, students are tested with the standardized Michigan Test of English Proficiency (Corrigan et al., 1979) and two in-house assessments to be placed into instruction levels. The in-house assessments measure listening and writing skills. The listening placement test score was chosen as the best measurement of initial proficiency (and treated as an independent variable). In order to choose the most reliable placement test score as the measure of initial proficiency for this study, correlations were run on each of the following: scores on the subsections of the Michigan test, the overall Michigan score, the in-house listening score, the in-house writing score, and the instruction level that the student was placed in. The correlation analysis indicated that the in-house listening test was most predictive of placement into instruction levels, $r = .838$ ($p < .001$).

Further support for the use of the listening test to represent initial proficiency comes from a construct validation study in which Sawaki, Stricker, and Oranje (2009) reported that the listening portion of the TOEFL Internet-based test loaded highly with general proficiency. Other researchers (e.g., Wang, 2009) have also used on listening-based placement tests to represent general proficiency across English L2 groups.

4.4 DATA

4.4.1 Data Collection

The Recorded Speaking Activity (RSA) is usually part of the speaking curriculum in English Language Institute at the University of Pittsburgh, but the number of RSA activities differs semester to semester. Typically, there is one (ungraded) introduction to the activity and two or three graded RSAs each semester. The introduction to the activity was not included for the present study. The RSAs are considered a valid speaking task for the curriculum, as they give the students an opportunity to speak on a general topic during the recording step, an opportunity to hear their own language during the transcription step of the activity, and an opportunity to notice a difference between the language produced and their explicit knowledge of the target language during the corrections step of the activity. This study only analyzes the initial recorded speeches, not the other steps of the activity. The teachers use the RSA to assess the students and to give individualized feedback to the students. The teachers grade each speech with an analytic grading rubric that includes elements of fluency, accuracy, grammatical and lexical complexity. Therefore, the task does not explicitly encourage the students to prioritize one of the components in language production. The students, however, tend to focus on grammatical accuracy over fluency, lexical variety, or grammatical complexity when completing the self-correction step of the task (McCormick & Vercellotti, 2009).

The week before the RSA task, three possible topics are reviewed and discussed in the classroom, but the chosen topic for the RSA is not known to the students until the task begins. After hearing and reading the topic, the students have one minute to plan how to address the topic. With pedagogical motivations, students are frequently reminded of the target academic

vocabulary words of the week as an encouragement to use them. The core vocabulary list is predetermined, and the words reviewed during the RSA session are not necessarily helpful for the given topic, especially since most topics are not of an academic nature (see description below). Since this research is not specifically addressing the use of individual words or types of words (like academic words), the encouraged words are not relevant here. It is mentioned here to point out there is a focus on vocabulary in the curriculum.

During the planning time, the students cannot take notes or use reference materials. The students are not specifically encouraged to monitor and to reformulate their language during the speech. Overall, the RSA activity is most similar to "no pre-task planning" groups described in Section 3.1.1 in the literature review. As such, the students' language performance reflects pressured online planning. Importantly, all RSAs are similarly administered. The speeches were recorded during regular speaking class time in a language media lab on Apple Power Mac computers with software developed with the Revolution Studio 2.6.1 package (Shafer, 2006). Since the RSA speeches are recorded in a computer lab in response to a given topic, they are not naturally occurring language samples. The task, however, is more representative of natural language performance than other experimentally elicited samples.

Each RSA has a different topic, and each student has flexibility in how to address the topic, but the students are not free to choose their topic. The topics (Table 8) differ from semester to semester and by instruction level. Since the population varies semester to semester and by instruction level, the number of speeches per prompt also varies. The teachers choose the RSA topics based on interest and based on appropriateness in relation to the syllabus (i.e., usually, one topic per semester is expected to elicit the past tense. An example topic from the low-intermediate level was "Describe your best friend from childhood. How did you meet? What

qualities help describe your friend? What did you used to do together?" A topic from the high-intermediate level was "Describe a custom, in your culture or another culture, which you do not like. Give details about the expectations of the custom, and describe the things that you don't like about it". A full listing of the topic prompts is given in Appendix A. Generally, the topics have the feature "familiarity of information" meaning that the topic allows the students to speak about a topic that would be familiar, often personal. It might seem as though language performance on familiar and personal topics would enhance performance particularly for fluency or accuracy, because personal information is easier to access which frees attentional resources. Familiar information, however, has not been shown to guarantee that the speaker will have more resources available to attend to the language performance (Skehan, 2001). Importantly, Skehan and Foster (1997) report similar fluency and accuracy means on the personal information task and the decision making task. In the current study, all of the RSA tasks are information-related talk, but the topics chosen seem to shift from factually oriented talk (description or narration) to evaluative talk (opinion-expression or evalution) (Luoma, 2004) at the highest level, which seems comparable to Skehan and Foster's personal-information task and decision-making task. For instance, a topic from the advanced level was "When the gap between the rich and the poor is so large, you need to balance a desire for luxury with compassion for the needs of others. Do you agree or disagree? Why?" Considering this shift from familiar to evaluative talk, it is noted that in English L2 writing research argumentative essays tend to have higher gramatical complexity than narratives (Lu, 2011).

Although the multiple topics increases variability, one advantage of the different topics is that any differences in initial scores or different change trajectories in the language performance will not be solely the effect of topic. For instance, if all students had the same first topic, initial

scores could be influenced by that specific topic; and if all students had the same sequence of topics, any change (increase or decrease) could be driven by (unknown and unplanned) differences in the topic.

Table 8 RSA Topics by Instruction Cohort and Level

cohort 1			cohort 2	
low-intermediate	high-intermediate	advanced	low-intermediate	high-intermediate
childhood meal (n = 10)	world problem (n = 24)	media violence (n = 16)	best friend (n = 16)	ideal vacation[4] (n = 1)
transportation (n = 10)	a regret (n = 24)	computerized society (n = 6)	a surprise (n = 16)	renting (n = 1)
admired person (n = 10)		internet risks (n = 8)		home city (n = 42)
		extravagant lifestyle (n = 5)		ideal job (n = 13)
		rich and poor (n = 8)		disliked custom (n = 42)
				famous person (n = 42)

4.4.2 Data Transcription and Coding

Since learners' oral data are challenging to analyze, reliability of the measures is threatened (Ellis & Barkhuizen, 2005, p. 163). In response, all measures are fully documented here. The speeches were transcribed using PRAAT software by a native speaker of English, who was trained and experienced in transcribing non-native speech. The author (also a native English speaker trained and experienced in transcribing non-native speech) then checked each transcription for accuracy and coded the data into clauses and AS units (Foster, Tonkyn, and Wigglesworth, 2000) which are sentence length utterances defined for oral language. A single

[4] There are single observations of some topics because some students returned to the same proficiency level instead of promoting to the next level with their peers and were given different topics than the promoted students.

modification was made when applying this measure. Sometimes learners did not produce a copula in utterances that were otherwise clearly an AS unit. For instance, this utterance was coded as an AS unit:

[1118] (2.162) *he in varied ah (1.405) differen(t) roles (9.429)* [^c].

Utterances with a missing copula are not specifically included as having AS unit status by Foster et al. (2000) because they do not have a finite verb, they do not fit the definition of independent sub-clausal since they are not elided versions of independent clauses, and they do not fit the description of minor utterances (such as *thank you very much*). However, these utterances function as an AS unit in the speech and have more meaning and complexity than a minor utterance, even without the copula. In addition, if these copula-missing utterances were not considered a separate AS units, the word counts associated with these utterances would artificially inflate the subsequent AS units. As such, these utterances were coded as AS units.

Using CLAN (Computerized Language Analysis) tools (MacWhinney, 2000), the language produced in each speech was coded and counted using available commands, such as mean number of words per utterance (which was coded to equal AS units) and part of speech tagger, in order to calculate the dependent variables described below.

4.4.3 Dependent Variables

The three components of language performance each have multiple subcomponents, necessitating more than one measurement of each component. The chosen measures (Table 9) were considered to be the most appropriate for the data in the study, considering the existing literature.

Table 9 Summary of Measurements for Each Speech

construct	measurement	defined as	subtype
complexity	length of AS unit (C1)	mean number of words / AS unit	global
complexity	clause length (C2)	mean number of words/clause	subclausal
complexity	clauses/AS unit (C3)	ratio of clauses/AS unit	subordination
complexity	lexical variety (C4)	D score	lexical
accuracy	AS unit accuracy (A1)	proportion of error free AS units	AS unit level
accuracy	clause accuracy (A2)	proportion of error free clauses	clause-level
fluency	phonation time ratio (F1)	speaking time (excluding filled pauses) divided by total time	global
fluency	mean length of pause (F2)	average length of (filled and unfilled) pause (>200 ms)	fluency breakdown
fluency	mean length of fluent run (F3)	average number of syllables in utterance bounded by pauses >200 ms	fluency proceduralization

4.4.3.1 Accuracy Measures

In this research of language performance, the focus was on learners' ability to produce accurate language. As such, accuracy was measured as percentage of error-free units at two levels of production: proportion of error-free clauses following Skehan (2001), Ellis and Barkhuizen (2005, p. 151), and Skehan and Foster (1997) and percentage of error-free AS units. Only absolute errors, rather than dispreffered forms (Ellis & Barkhuizen, 2005, p. 59) based on standard American English were counted as errors. Errors in syntax, morphology, and lexical choices were coded; pronunciation errors were not. Whenever the speech had a self-correction, only the final version was considered. Thus, when a student made an accurate self-correction, that unit was considered error-free, again following Ellis and Barkhuizen (2005, p. 49). Following Schachter and Celce-Murcia (1977) both of these accuracy measures represent relative frequency of errors (or lack thereof) rather than raw numbers and focus on what the participants can do. Both are general or global measurements of accuracy, albeit on two levels of production, which is better suited for the longitudinal data than a specific measure (e.g., article

suppliance, subject-verb agreement) based on the review of the measures in Chapter 2 Section 1.1 and the research questions. Specifically, the literature does not address trade-off effects at a more specific level of accuracy (i.e., there are no predictions that an increase in grammatical complexity, lexical variety, or fluency will affect accuracy of any specific forms). Trade-off effects are only at a general level of accuracy, in that specific student errors cannot be predicted. If fluency is prioritized, accuracy in general is at risk, not specifically, considering not all learners will be prone to make errors on the same specific structure.

4.4.3.2 Complexity Measures

The AS unit (Foster et al., 2000) was determined to be the most appropriate base unit for the data. Following Norris and Ortega's (2009) suggestion, each speech was measured for syntactic complexity at three different levels. Length of AS unit and clause length were measured in words. Subordination was calculated as clauses (finite and non-finite clauses) divided by AS units. Following Skehan's (2009) suggestion, each speech was also measured for lexical variety. Lexical variety was calculated using D (Malvern & Richards, 1997) based on word rather than lemma. (A comparison of the D scores based on word and based on lemma in these data found an extremely high correlation, r = .975, which indicates that using a lemmatized D score would have produced very similar results.)

4.4.3.3 Fluency Measures

Fluency was measured by phonation time ratio (PTR) as a general measure, by fluency breakdown with mean length of pause (MLP), and by fluency proceduralization with mean length of fluent run (MLFR). These three measurements have been used on similar data (De Jong & Perfetti, 2011). Phonation time ratio was calculated as the time speaking (excluding filled

pauses) divided by total time. Following De Jong and Perfetti, pause length was calculated as the average length of pauses of at least 200 milliseconds, including both filled (e.g., “uh”) and silent pauses. The mean length of fluent run was measured in the number of syllables uttered together bounded by pauses of at least 200 milliseconds.

4.5 DATA ANALYSIS

Descriptive-quantitative longitudinal research of second language performance has typically studied only small homogenous groups (or even single participant case studies) which warrants only descriptive statistics, not inferential statistics (Ortega & Iberri-Shea, 2005). This research’s scope enables both descriptive and inferential statistics.

4.5.1 Quantifiable Variables Study Design

In this research, each participant (n = 66) was measured repeatedly, typically one month apart approximately. The number of observations (speeches) and the spacing of the observations differ among participants because of timing constraints in the schedule within the semester and time between academic semesters. Time was treated as a random variable (fraction of year since enrollment) so that each measurement is associated with a specific point in the student’s learning history. Time was adjusted by approximately one month (.833 from fraction of the year since the start of semester) so that the intercept is approximately at the start of data collection. There was a minimum of three observations and a maximum of seven observations (speeches) per student, with an average of nearly five speeches for each participant. These observations, which are

called level-1, are the basis of each participant's individual growth trajectory (Raudenbush & Bryk, 2002, p. 161).

Each speech was scored for each of the dependent variables: two accuracy measures (percentage of error-free AS units and percentage of error-free clauses), three grammatical complexity measures (length of AS unit, clause length, clauses per AS unit), the lexical variety D-score, and three fluency measurements (phonation time ratio, mean length of pause, and mean length of fluent run).

Each speech was also coded with potentially predictive independent variables of age, gender, initial proficiency (measured by the listening placement test), language background (Arabic and non-Arabic), and instruction cohort (Table 10). These demographic variables are time-invariant variables, meaning that they retain the same value over all observations from a participant. Following recommendations from Singer and Willett (2003), age and initial proficiency were grand mean centered based on the person-level data, which means each observation was coded as its distance from the mean. Centering on person-level data, where the mean is calculated with one value for each participant, is important for these data because if the means were calculated on the value for each observation, the values of participants with more observations would skew the mean (i.e., the age of a student with seven observations would be included seven times while the age of a student with three observations would be included three times). Categorical time-invariant predictors (gender, L1, and cohort) were not centered, meaning that the values were not based on the difference from the (theoretical) mean. By keeping the categorical value, the results are easier to interpret because the basic value will represent an actual student (Arabic males in cohort 1) with a zero value for those variables,

rather than a hypothetical student with an "average value" for that variable (e.g., for the variable gender: .48).

Table 10 Summary of Independent and Dependent Variables

independent variables		dependent variables (also possible time-varying predictors)
time-invariant	time-varying	
gender	topic	A1 score
age (centered)		A2 score
initial proficiency (centered)		C1 score
language background		C2 score
instruction cohort		C3 score
		C4 score
		F1 score
		F2 score
		F3 score

4.5.2 Hierarchical Linear Modeling

4.5.2.1 Nonparametric

For each measure, the observed scores for each participant are presented as a set of ordinary least squares (OLS) trajectories, where each participant's data can vary. These trajectories are presented in a figure smoothed but not forced to fit a specific form (e.g., a straight line, a quadratic curve), which gives a truer picture of each participant's change trajectory. Since the smooth nonparametric OLS trajectories are not fitted trajectories, they cannot give numeric summaries of the trajectories. Also, random fluctuations from measurement error make nonparametric trajectory patterns less reliable for studying development (Singer & Willett, 2003). Therefore, after reviewing the data as smooth nonparametric OLS trajectories (presented in a figure), the data were fitted into a common functional form, described in the next section.

4.5.2.2 Parametric

Parametric models fit the each participant's regression model onto a common functional form (e.g., linear, quadratic), which allows further exploration of the data (Singer & Willett, 2003). The data were fitted using full-maximum likelihood hierarchical linear modeling (HLM). By choosing the same functional form across all participants, the researcher can use numerical data to differentiate participants or groups of participants (Singer & Willett, 2003, p. 28). The first form selected is a simple linear model, which fits the data to a straight line.

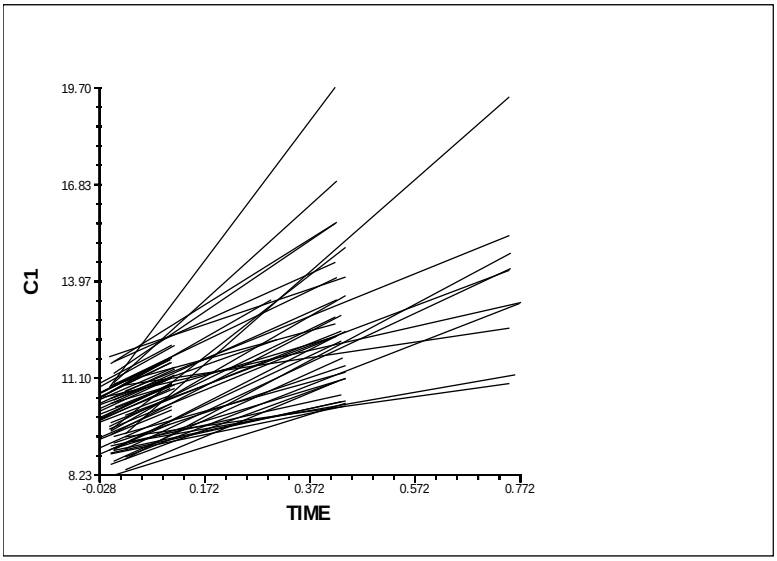

Figure 1 Length of AS unit (C1) Scores Fit to a Simple Linear Model

Figure 1shows the length of AS unit (C1) scores for all 66 participants fitted to a linear model. Then, for each measure the data were also fitted to a quadratic pattern because growth may not be neatly linear. A chi-square test of deviance statistics was run to determine whether the linear form or the non-linear trajectory better fit the data. The result of this test is reported for each measure. When deviance statistics supported a non-linear trajectory, the model building was

based on the non-linear unconditioned model. In other words, any predictors were added to the non-linear unconditioned model.

Each participant's initial score (intercept) and change trajectory (slope) for each measure was fitted to check for individual change over time. This step addressed my first research question, "Is there significant individual growth in learners' performance over time?" A change over time is required in order to study change, obviously.

After determining that the average slope was significantly different from zero, each student's language performance development was represented by the change (either growth or decline) trajectory, which includs the intercept and slope in the unconditional growth model (with time as the only predictor). The mean initial score and standard deviation of each measure (presented in a table) gives an indication of how much variation exists in the data for that measure at the first observation. Similarly, the mean slope and standard deviation reveals how much variation exists in the change trajectory among participants for that measure.

4.5.2.3 Conditioned (with Predictors) Models

If sufficient variation exists in the data, more analysis is warranted. Whenever the unconditioned growth model showed significant variance in the data either in the initial scores or in the slope, predictors were added to the model. Whenever the random variance component indicated that there was no significant variation, however, the random variance component was constrained to zero.

When there was significant variation in the initial values for each measure, first, time-invariant level-2 person-level predictors (i.e., independent variables based on demographic information) were added to the model in an attempt to explain the variance. This step addressed the second research question regarding individual differences (demographic) predictors related to

growth that might explain individual growth, using the variation in initial scores or change trajectories.

Differences in the rate of growth are necessary to test which variables affect growth (Boyle & Willms, 2001). When there was variation in the rate of growth (slope) among learners, again, time-invariant person-level predictors (i.e., independent variables) were added to the model in an attempt to explain the difference. The time-invariant variables included: gender, age, initial proficiency, language background, and cohort.

In order to assess whether a variable in the intercept (initial score) or slope (change rate) model is significantly predictive, models' residual variances, measured by Akaike Information Criterion (AIC), (Singer & Willett, 2003, pp. 121-122), were used to compare models. All subsequent conditioned models were compared to the unconditional growth model in order to evaluate if the added predictive variables significantly improved the fit of the model. In order to determine how much of the variance is explained by the subsequent conditioned models, the researcher can compare the amount of variance in the unconditioned model to the remaining variance in the conditioned model. This allows a calculation of the amount (percentage) of the variance explained. Although this measure should not be strictly understood as effect size, this measure has been used as a measure of effect size in HLM (Roberts & Monaco, 2009).

When variation remained after considering each and every time-invariant predictor, time-varying predictors (predictors whose values may vary observation to observation) were added to the model when there was an a priori directional relationship between the variables. The model then controlled for effect of the time-varying predictor on the dependent variable, like a covariant in ANCOVA. As such it could explain antecedent-consequent relationships (Ortega & Iberri-Shea, 2005).

In sum, HLM analysis was used to find the best-fitting model to the data (presented in a table) to investigate the change (rather than status) in language performance. The entire data set, the observed score for each of the nine measures of each speech (n = 294) ordered by participant (n = 66), is found in Appendix C.

4.5.2.4 Rationale for Using Hierarchical Linear Modeling

HLM can model the form and the predicting factors of individual growth (Raudenbush & Bryk, 2002, p. 161). HLM allowed for a better analysis of the current data since it uses all available data, rather than limiting the analysis to participants with a full data set (Singer & Willett, 2003) and because it does not aggregate data.

First, HLM is especially useful for analyzing longitudinal data as it models observations from individuals. Longitudinal studies are often hindered by attrition, but HLM does not require an equal number of observations. This advantage of using all available data is especially relevant because as described in the Methodology section, the participants differed in the number of observations (speeches). Also, since each observation is coded by the date of the speech, the model can consider the time between the observations, rather than assuming that subsequent observations are equally spaced, which is especially important because observations within an academic semester can be closer than observations between semesters.

Second, researchers often aggregate all data, which increases sample size and power. But findings based on aggregated data must then only describe the class, not the individuals. And, when aggregating data over repeated measures, the results give group trends while ignoring individual trajectories. HLM allows an exploration of individual trajectories, on a more complex level than simple correlations or group mean comparisons. Importantly, HLM allows intercepts and slopes to vary across participants (Singer, 1998), which allows the model to explain more of

the variability. This analysis technique can indicate predictive variables for outcome variables (performance measures) and for rate of growth which is relevant when testing a dynamic system theory approach to research. Thus, this methodology has two important strengths, it can describe the complexity, accuracy, and fluency of the language performances and describe their relationships overtime. Therefore, these data can address Skehan's (2009b) concern that different students must be driving the growth in multiple components when trade-off effects are not found.

Further, HLM also can model linear and nonlinear growth models, which is valuable in developmental studies in which nonlinear models are plausible, such as, language performance.

This type of modeling has been used in developmental psychology and in educational psychology, frequently in longitudinal studies of child development. HLM has been used to study the impact of early approaches to learning on academic performance (Li-Grining et al., 2010), to determine if reading disabilities are deficits or delayed learning (Francis et al., 1996) to compare reading development in native English and English language learners (Lesaux, Rupp, & Siegel, 2007), to determine early predictors of biliteracy development (Jared et al., 2010). HLM is not limited to child development longitudinal studies; it has also been used to analyze longitudinal data of college students' performance and growth (Strauss & Volkwein, 2002) and to investigate gender (in)equality in college faculty positions (Umbach, 2007), for example.

4.5.3 Correlations Analysis

My third research question asked, "What are the relationships between the CAF language performance measures?" In order to more fully answer the third research question, the pooled within-individual and between-individual correlations of measures were calculated on the full data set. An interclass correlation coefficient (such as Pearson *r*) is the common way to measure

the relationship between two variables (McGraw & Wong, 1996). However, this standard test requires the data to meet the assumption of independence between observations (Hox, 2010, p. 4). Since this research included multiple observations of participants, these data violate that assumption. It is expected that observations from the same individual will be more correlated, obviously, than observations from different individuals. In addition, the correlations must be adjusted to account for the uneven number of observations from each participant.

An important benefit of multiple-level (observations of individuals and individuals in groups) analysis is that the total variance can be separated into the within-individual and the between-individual components (Van de Pol & Verhulst, 2006). The *intra*class correlation coefficient (ICC) measures the proportion of variance in the outcome that is found at level-2 (Raudenbush & Bryk, 2002, p. 36), which is between participants in this study. The intraclass correlation, or the cluster effect, measures how consistent measures within a group are. In other words, the intraclass correlation is the proportion of the variance between participants compared to the total variance in the population (Hox, 2010, p. 15)

$$\text{ICC} = \frac{\text{group level (between participants) variance}}{\text{total variance}}$$

4.5.3.1 Between-individual Correlations

Between-individual correlations state the strength of the relationship between two variables in the data. In this dissertation, I report the regular between-individual covariance matrix because the estimated between-individual covariance matrix over-adjusted as a result of standard deviations close to zero (particularly with the complexity measures).

Most importantly, correlations at the group level should not be used to make inferences about the individual (Ostroff, 1993). This means that if two measures are correlated in

aggregated group data, it does not necessarily mean that those measures are correlated at the individual level. In addition, aggregating data tends to inflate correlations (Ostroff, 1993), which could lead to apparently significant correlations that would not be found in non-aggregated data.

4.5.3.2 Within-individual Correlations

Within-individual correlations can be used when measuring students repeated. These correlations tell how much you can predict a participant's score on a measure when you know his score on the other measure. Within-individual correlations are more valuable when considering possible trade-off effects within the individual's language development, since any trade-off effects would occur within an individual trying to produce language. Again, only within-individual correlations are really of interest since between-individual correlations could be driven by different influences.

4.5.3.3 Interpretation of Correlations

When the measures within a construct (accuracy, complexity, or fluency) are strongly correlated, the finding suggests that the measures are tapping the same construct and, therefore, the multiple measurements might not be needed. When measures within a construct are not highly correlated, it implies that the measures are tapping different aspects of the construct, and multiple measures are therefore useful. If measures across components are positively correlated, the findings suggest that the components are "connected growers". If measures across components are negatively correlated, the findings could suggest there are trade-off effects between the components. When the correlations are close to zero, the findings suggest the components are not consistently related.

4.6 SUMMARY OF METHODOLOGY

This chapter reports on the method of a longitudinal which study analyzes observations (n = 294) from 66 participants over time in the English Language Institute at the University of Pittsburgh during the three academic semesters of 2010. The participants were male (n = 34) and female (n = 32) young adult learners (18-35 years, mean 25.3 years) of English from three language backgrounds (Arabic, Chinese, and Korean) from two instruction cohorts.

There were multiple observations (3-7) from each participant, recorded over three to nine months. The average number of observations was 4.45, with 62.1% of the participants supplying four or five speeches, which fairly represents the average length of enrollment. Each observation was a two-minute recorded monologue (RSA) on a given topic. The number of RSAs each semester differed, and the topics differed for each RSA. Each speech was coded to measure the complexity, accuracy, and fluency of the oral performance. The coded dependent variables were length of AS unit (C1), clause length (C2), clauses per AS unit (C3), lexical diversity (C4), percentage of error-free AS units (A1), percentage of error-free clauses (A2), phonation time ratio (F1), mean length of pause (F2), and mean length of fluent run (F3). Length of AS unit (C1) and length of clause (C2) were measured in words. Clauses included finite clauses and non-finite clauses with an adjunct or complement. Lexical diversity was measured by D. Phonation time ratio is the time spent speaking (excluding filled pauses) divided by total time. Pauses were defined as filled or unfilled pauses lasting 200 milliseconds or more. Fluent runs are the stretch of speech, in syllables, between pauses of 200 milliseconds or more. These nine measures were expected to capture the multiple components of second language development.

Since the focus of the study language development or change (as opposed to language performance status), the data were analyzed using hierarchical linear modeling (HLM). The

observations were nested within individuals (level-2). The time-invariant independent variables (predictor variables in the model) included initial proficiency (grand mean centered), age (grand mean centered), gender, language background, and instruction cohort. Time-varying predictors included time in the program (adjusted to the start of data collection), topic, and the other outcome variables (the other dependent variables.) HLM can model both linear and non-linear change trajectories and can evaluate how useful predictor variables are in explaining the variance in initial scores and the variance in change trajectories. Importantly, HLM does not require an equal number of observations per participant, and it does not require spacing between observations. The relationships between the dependent variables were also analyzed with within-individual correlations (which shows how the score on one measure correlates with another's within individuals) and between-individual correlations (which shows relationships at the group level).

5.0 RESULTS

The results of the study are given first by construct: accuracy (Section 5.1), complexity (Section 5.2), and fluency (Section 5.3). Within each construct, the results of each measure are given, in the following order: the nonparametric figure of the data, the mean initial score and mean slope for the measure, the unconditioned parametric growth model, the best-fitting parametric model with time-invariant predictors (if necessary), the best-fitting parametric model with time-invariant and time-varying predictors (if necessary). The results from the unconditioned parametric growth model answer my first research question, "Is there significant individual growth in learners' performance over time?" The parametric models with time-invariant predictors answer my second research question, "Do individual differences explain individual growth?" The demographic information (time-invariant independent variables) include gender (as a categorical variable), language background (as a categorical variable), instruction cohort (as a categorical variable), age (grand mean centered) and initial proficiency (grand mean centered) based on the listening placement test given at enrollment. My third research question was, "What are the relationships between the CAF language performance measures?" The within construct correlations begin to answer this research question for the relationships within the construct. These correlation results are presented after the results of each measure in the construct. The between construct facet of this question is partially answered by any parametric models with time-varying predictors and by between CAF construct correlations. To fully answer the third

research question, the final section (Section 5.4) describes the correlations found in the data between the CAF constructs. At the end of each result section, a quick summary and discussion is offered.

5.1 ACCURACY

Accuracy was measured as the percentage of error-free AS units (A1) and error-free clauses (A2). In general, I hypothesized improvement in all measures. I expected initial proficiency to influence initial scores (i.e., participants with higher initial proficiency would have higher initial scores) and growth rate (i.e., participants with lower initial proficiency would have a greater change rate). I did not expect differences based on age, gender, or language background, or instruction cohort.

5.1.1 Percentage of Error-free AS units (A1)

Figure 2 shows the collection of smooth nonparametric and ordinary least squares (OLS) trajectories across participants (n = 66) for percentage of error-free AS units (A1). Time was adjusted so that the intercept is approximately at the start of data collection. The linear function was a better model for the data; the possibility of a non-linear trajectory was rejected, χ^2 (4) = .758, p = .944.

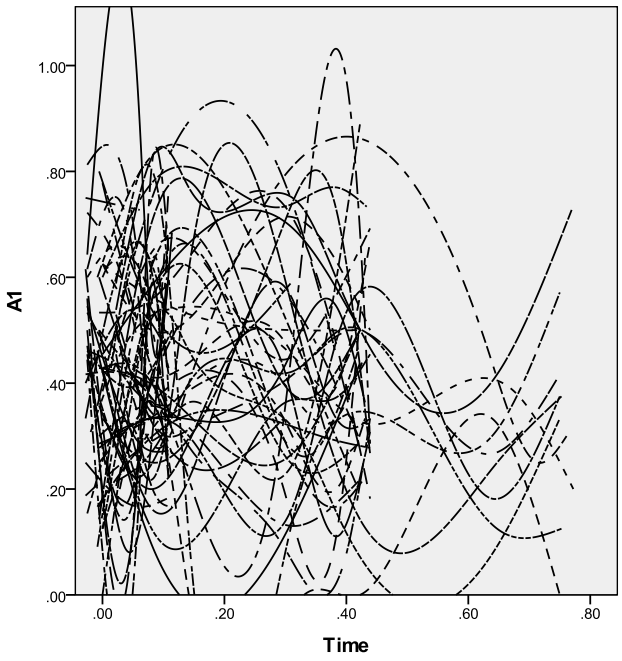

Figure 2 Collection of Smooth Nonparametric and OLS Trajectories of A1 Scores

Reviewing the nonparametric data in Figure 2, the initial scores seem to be between 20% and 50% accuracy at the AS unit level. There was no clear increase or decrease trajectory, overall, but the data show some large variation among and within individuals. The parametric linear model descriptive statistics (Table 11) clarify that the average estimated initial score was .436 (which can also be found listed as the coefficient for initial status in Table 12) with a standard deviation of .101. The average estimated slope was -.105 (which is also found under coefficient of mean growth rate in Table 12) with a standard deviation of .013.

Table 11 Descriptive Statistics for Individual Growth Parameters of A1 (n=66)

	Initial status (intercept)	Rate of change (slope)
Mean	.436	-.105
Standard deviation	.101	.013

This finding means that the average-aged (25.3 years) and average proficiency (19.2 on the placement test) student in this study had an observed percentage of error-free AS units (A1) score of 43.6% (SD = 10.1%) when entering the program and that his score *decreases* by an estimated 10.5% (SD = 1.3%) over the year. The magnitude of the standard deviations suggests that the participants differ considerably in their fitted initial scores, but not in their change rate. The intraclass correlation for this measure was .3314. In other words, the variance explained by between-individual differences (group-level) is 33.1%, which warrants further analysis.

Table 12 Unconditioned Linear Model of Growth of Percentage of Error-free AS units (A1)

Fixed Effects	Coefficient	SE	*t*	*p*
Mean Initial status, B_{00}	.436	.017	25.79	<.001
Mean growth rate, B_{10}	-.105	.039	-2.67	.010
Random Effects	Variance Component	*df*	X^2	*p*
Initial status, r_{01}	.010	65	114.809	<.001
Change rate, r_{1i}	.0002	65	49.184	>.500
Level-1 error, e_{ti}	.023			

The full results of the unconditioned linear growth model (Table 12) lists the expected A1 score for the average-aged and average proficiency student to be .436 at one month in the program. The mean linear growth rate of change was estimated to be -.105 per year, indicating an average rate of decrease of AS unit accuracy during the study. Both the mean intercept and slope were statistically significant, indicating that both parameters are necessary for describing the mean change trajectory. Students varied significantly in their A1 accuracy at one month ($\chi^2_{65} = 114.809$, $p < .001$) but did not vary significantly in their change rate ($\chi^2_{65} = 49.184$, $p > .500$).

Time-invariant predictors (i.e., independent variables) were added to the model to explain the variance in initial scores, which is warranted by the significant *p*-value of the initial status variance component. Since there was no significant difference in slopes (shown by the > .500 *p*-

value of the change rate variance component), the random variance component was constrained to zero, and no variables would be relevant for predicting the change trajectories (slope).

Table 13 Conditioned Linear Growth Model of Percentage of Error-free AS units (A1)

Fixed Effects	Coefficient	SE	*t*	*p*
Model for initial status, π_0				
Intercept, β_{00}	.457	.020	22.57	<.001
L1, β_{01}	-.062	.029	-2.12	.038
Initial Proficiency, β_{02}	.006	.003	2.04	.046
Model for Growth Rate, π_1				
Intercept, β_{10}	-.100	.039	--2.59	.010
Random Effects	Variance Component	*df*	X^2	*p*
Initial status, r_{01}	.0095	63	190.273	<.001
Level-1 error, e_{ti}	.023			

The best linear growth model (Table 13) included initial proficiency and L1 as level-2 predictors to explain the variation in the initial scores. The expected A1 score for average-aged and average proficiency Arabic students at one month was estimated to be 45.7%. Initial proficiency was strongly related to A1 at one month. On average, for every one point increase in (centered) initial proficiency, there was a .6% increase in A1 accuracy scores. Language background was strongly related to A1 at one month. On average, non-Arabic students had lower A1 scores by 6.2% at one month. All three parameters (initial scores, L1, and initial proficiency) were significantly different from zero, indicating that all parameters are necessary for describing the mean growth trajectory. As stated earlier, there was no variation in growth rate between students; on average, all students had a decrease in A1 scores by 10% per year in the program. Students still varied significantly in their A1 accuracy at one month ($\chi^2_{64} = 190.279$, $p < .001$). By comparing the variance component of the initial score, the proportion of intercept variance explained by initial proficiency and language background was calculated to be 5.0%.

The HLM equation of the best fitting model with time-invariant predictors shows that average Arabic L1 students had higher initial A1 scores than non-Arabic L1 students. In addition, for every increase in initial proficiency score, there was a corresponding increase in accuracy. The change rate did not differ; all groups show a decline in A1 accuracy.

Table 14 Conditioned Linear Growth Model of A1 with Time-Varying Covariate C1

Fixed Effects	Coefficient	SE	*t*	*p*
Model for initial status,				
Intercept, β_{00}	.644	.044	14.81	<.001
L1, β_{01}	-.080	.029	-2.74	.008
Initial Proficiency, β_{02}	.012	.003	3.62	.001
Mean Growth rate, β_{10}	.027	.044	.616	.539
Mean C1 slope	-.018	.004	-4.63	<.001
Random Effects	Variance Component	*df*	X^2	*p*
Initial status, r_{01}	.002	63	56.272	>.500
C1 slope, r_2	.00003	65	62.658	>.500
Level-1 error, e_{ti}	.021			

The overall percentage of error-free AS units (A1) scores seem disappointing in that the accuracy scores are generally lower after time in the program. Considering the complex nature of language performance, however, this finding would be more disappointing if there was not growth in other areas. Since longer AS units are mathematically less likely to be error-free, the A1 scores were analyzed while controlling for length of AS unit (C1). Adding this time-varying predictor to the model, explained all of the remaining variation.

Results of the best fitting linear growth model with the addition of the time-varying covariate (Table 14) specify that when average-aged and average proficiency Arabic L1 students produce average length AS units (C1), the expected A1 accuracy score was estimated to be 64.4%. Non-Arabic L1 students had a lower initial A1 score by 8.0%. For every increase of one unit (centered) initial proficiency score, the percentage of error-free AS unit score is expected to

increase by 1.2%. The mean linear growth rate for all students was estimated to be 2.7% (now an increase) while controlling for length of AS unit (C1) scores. At one certain time point, a one word increase in AS unit length (C1) decreased the A1 scores by 1.8%. All parameters (except for growth rate) were significantly different from zero, indicating that all variables (initial proficiency, language background, and length of AS unit) are necessary for describing the mean growth trajectory. After controlling for length of AS unit (C1), students no longer vary significantly in their AS unit level accuracy (A1) at one month in the program ($\chi^2_{63} = 56.272$, $p > .500$), or in the relationship between percentage of error-free AS units (A1) and length of AS unit (C1) scores at a time point ($\chi^2_{65} = 62.658$, $p > .500$).

After controlling for length of the AS unit (C1), again, Arabic students had higher percentage of error-free AS unit (A1) scores than non-Arabic students. For every one unit increase in centered initial proficiency, there is an expected 1.2% increase in percentage of error-free AS units (A1) score. By controlling for length of AS unit, the data showed that accuracy increases by 2.7% for all groups over a year in the program.

5.1.2 Percentage of Error-free Clauses (A2)

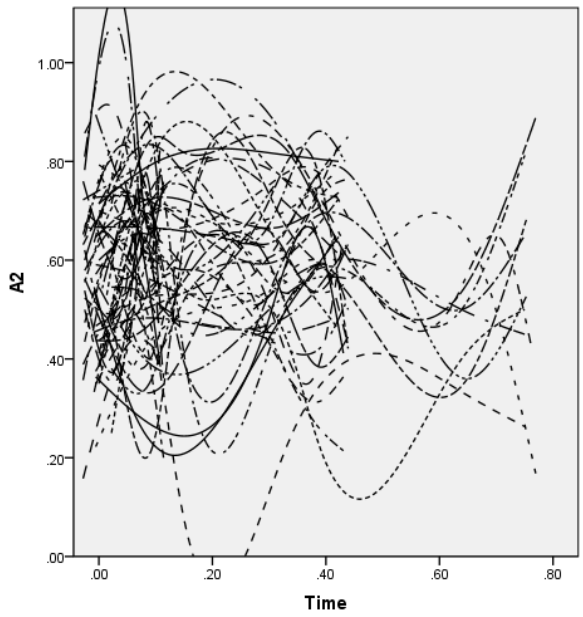

Figure 3 Collection of Smooth Nonparametric and OLS Trajectories of A2 Scores

Figure 3 illustrates the collection of smooth nonparametric and OLS trajectories across participants for percentage of error-free clause (A2) scores. Time was adjusted so that the intercept is approximately at the start of data collection. The possibility of a non-linear trajectory was investigated and rejected, χ^2 (4) = 4.052, p = .405. The parametric linear model descriptive statistics (Table 15) found an average estimated initial score across the data of .586 (SD = .101) and the average estimated slope of .030 (SD =.129), indicating a small average rate of increase of A2 accuracy (percentage of error-free clauses) during the study.

Table 15 Descriptive Statistics for Individual growth Parameters of A2 (n=66)

	Initial status (intercept)	Rate of change (slope)
Mean	.586	.030
Standard deviation	.101	.129

The magnitude of the standard deviation in initial scores suggests that the participants differ considerable in their fitted initial scores. In fact, the intraclass correlation was .4159, which indicates the data cluster at the between-individual level. Results of the unconditional growth model (Table 16) specify that the mean intercept was statistically significant, indicating that only this parameter is necessary for describing the mean growth trajectory. Students varied significantly in their A2 accuracy at one month (χ^2_{65} = 159.806, p < .001) but not in their growth rate (χ^2_{65} = 65.028, p = .476). Since the variance component for change rate was not significant, it was constrained to zero.

Table 16 Unconditioned Linear Model of Growth in Percentage of Error-free Clauses (A2)

Fixed Effects	Coefficient	SE	*t*	*p*
Mean Initial status, B_{00}	.586	.016	36.77	<.001
Mean growth rate, B_{10}	.030	.044	.68	.497
Random Effects	Variance Component	*df*	X^2	*p*
Initial status, r_{01}	.010	65	159.806	<.001
Change rate, r_{1i}	.017	65	65.028	.476
Level-1 error, e_{ti}	.014			

Level-2 predictors were added to the intercept model to determine how much of the variation in the initial score can be explained by time-invariant (independent) variables. Initial proficiency and L1 were found to be the only time-invariant predictors in the model for percentage of error-free clauses (A2). Results of the linear growth model with initial proficiency and L1 as level-2 predictors for the intercept (Table 17) specify that the expected error-free clause (A2) scores of average-aged and average proficiency Arabic students was estimated to be .610 (61.0%). Language background and initial proficiency were strongly related to clause accuracy (A2) scores at one month. On average, non-Arabic L1 students had lower initial A2 scores by .068 (6.8%). For every one unit increase of centered initial proficiency, there was a

corresponding increase in percentage of error-free clause (A2) scores by .013 (1.3%). The mean intercept, L1, and initial proficiency were significantly different from zero, indicating that all parameters are necessary for describing the mean growth trajectory. The mean linear growth rate was .031 (3.1%). Students still varied significantly in their A2 scores at one month, after controlling for initial proficiency and language background (χ^2_{63} = 215.354, p <.001). By comparing the variance component of the initial status from Table 16 and Table 17, the proportion of initial score variance explained by initial proficiency and language background was found to be 7.5%.

Table 17 Conditioned Linear Growth Model of Percentage of Error-free Clauses (A2)

Fixed Effects	Coefficient	SE	*t*	*p*
Model for Initial status				
Intercept, β_{00}	.610	.016	37.26	<.001
L1, β_{01}	-.068	.025	-2.70	.009
Initial Proficiency, β_{02}	.013	.003	4.77	<.001
Mean growth rate, B_{10}	.031	.047	.65	.519
Random Effects	Variance Component	*df*	X^2	*p*
Initial status, r_{01}	.085	63	215.354	<.001
Level-1 error, e_{ti}	.015			

Since the variance component of the initial scores was still statistically significant, more variance could possibly be explained. The time-varying predictor of clause length (C2) was added to the model and was found to be a predictor of the A2 scores. Results of the linear growth model with time-varying covariate (Table 18) specify that for average Arabic L1 students with average clause length (C2) scores, the expected clause accuracy score was estimated to be .872 (87.2%). Non-Arabic L1 students had a lower initial score by .063 (6.3%). For every one unit increase in centered initial proficiency, there was .014 (1.4%) increase in percentage of error-free clauses. At one certain time point, one word increase in clause length (C2) scores decreased the

clause accuracy (A2) scores by .045 (4.5%). The mean linear growth rate for all students was estimated to be .084 (8.4%) while controlling for clause length (C2) scores. All parameters were significantly different from zero, indicating that all variables are necessary for describing the mean growth trajectory. After controlling for clause length (C2), students no longer varied significantly in their clause accuracy (A2) at one month in the program (χ^2_{63} = 60.335, p >.500), or in the relationship between clause accuracy (A2) and clause length (C2) scores at a time point (χ^2_{65} = 65.989, p =.443).

Table 18 Conditioned Linear Growth Model of Clause Accuracy (A2) with Covariate C2

Fixed Effects	Coefficient	SE	*t*	*p*
Model for initial status				
Intercept, β_{00}	.872	.046	18.89	<.001
L1, β_{01}	-.063	.024	-2.63	.011
Initial Proficiency, β_{02}	.014	.002	6.30	<.001
Mean Growth rate, β_{10}	.084	.038	2.22	.028
Mean C2 slope, β_{20}	-.045	.008	-5.83	<.001
Random Effects	Variance Component	*df*	X^2	*p*
Initial status, r_{01}	.001	63	60.335	>.500
C2 slope, r_2	.0003	65	65.989	.443
Level-1 error, e_{ti}	.012			

Considering all of the relevant predictors (language background, initial proficiency, and mean length of clause) of percentage of error-free clause (A2) scores, average Arabic students had higher A2 scores than non-Arabic students. For every word increase in clause length, there was a corresponding decrease in clause accuracy of 4.5%. For every one point increase in centered initial proficiency, there was a corresponding increase in clause accuracy of 1.4%. As stated above, the rate of growth did not differ among the students.

5.1.3 Correlations between the Accuracy Measures

Unsurprisingly, the two accuracy measures had strong positive correlations. The between-individual correlation was r = .876. More importantly, the pooled within-individual correlation (r =.700) was significant at the .01 level. This indicates that these measures were strongly correlated when measuring students repeatedly.

5.1.4 Discussion of Accuracy Results

At first blush, the overall accuracy numbers in the unconditioned growth models were surprisingly low, only 43.6% of AS units were error-free and 58.6% of clauses were error-free. On the other hand, perhaps these accuracy rates are not low, considering that Ahmadian (2011) reports error-free clause scores of between 38% and 39% for his population of intermediate EFL learners. In addition, Wang (2009) reported a similar range (35.5%-36.9%) in her study of undergraduate college students beginning to study in an ESL environment. While the A2 scores showed modest gains (3%) over the year in the program, A1 scores showed a decrease of 10.5% in accuracy over the year. This finding was especially disappointing since grammatical accuracy is an important goal of instructed SLA and since grammatical accuracy is the main focus of the students (McCormick & Vercellotti, 2009).

For both accuracy measures, initial proficiency, language background, and length of unit explained most of the variance in initial scores (Table 19). The fact that initial proficiency explains variation in accuracy scores is rather expected. As Pallotti (2009) points out, it is difficult to separate the construct of accuracy from proficiency.

The impact of language background on accuracy scores, however, was un expected. It was assumed that the use of general measures of accuracy and the use of a ratio of error-free clauses (rather than raw number of errors) would limit L1 effects. L1, however, was found to be a significant predictor (stronger even than initial proficiency and length of unit combined) both accuracy models. This finding of language background impacting general measures of language performance has not been previously reported. First, the previous research has often been limited to a single L1 population, which, obviously, cannot address the impact of different language backgrounds. Second, previous research with participants from multiple language backgrounds (e.g., Bardovi-Harlig & Bofman, 1989) were much smaller studies, with only six participants from each language background. Without further study of the data, there is no explanation for the finding that the average Arabic student had higher initial scores than the average non-Arabic student, even after considering differences in initial proficiency.

Table 19 Summary of Accuracy Measures Results

measure	difference in initial score?	difference in slope?	predictors
Percentage of Error-free AS units (A1)	yes	no	initial proficiency language background length of unit
Percentage of Error-free Clauses (A2)	yes	no	initial proficiency language background length of unit

For both accuracy measures, the length of the unit also impacted accuracy. Increases in length corresponded to a decrease in accuracy (i.e., longer units tended to have more errors than shorter units). This result is a function of the calculation of error-free units. Further, initial proficiency has a bigger impact and language background has bigger impact on accuracy scores,

at both levels, than length of unit. Therefore, the data do not necessarily support a trade-off hypothesis.

Further, the time-varying phonation time ratio was found not to be a significant predictor. This finding, from this population at this proficiency level, is contrary to the Higgs and Clifford theory that an emphasis on fluency negatively affects accuracy in the long run.

Turning to rate of change, neither accuracy measure showed significant variation in change trajectory. The lack of variance in change trajectory (slope) indicates that each group equally benefits from this program. This finding, however, is interesting because higher initial proficiency did not corresponded to steeper gains over the year of study, as found in reading ability (Stanovich, 1986). This finding might suggest that accuracy or oral performance cannot be sped up.

There is much variability in the accuracy scores over time, especially within individuals, which supports dynamic systems theory of development. Contrary to dynamic systems theory, non-linear trajectories of accuracy were not a better fit for the data. Also, in this study with 66 participants, there is no evidence of clustering of the variation to support Larsen-Freeman's (2006) proposal that students take different paths through development. In addition, there was no evidence that students have different trajectories, based on initial proficiency, L1, age, gender, or instruction cohort.

Measuring accuracy at two levels was useful in that the two measures give a more complete picture of the accuracy percentage norm for participants at this proficiency level and give a more complete picture of the development of the subsystems, including in relation to the development of grammatical complexity. The correlations between the two measures of accuracy are quite high. This finding might allow researchers to use only one of the measures, depending

on the level of analysis. If only one measure is chosen (for practical reasons), however, clause level accuracy is recommended. Although the development of accuracy at both levels is important, producing error-free AS units was difficult for this population. Also, these results suggest that clause level accuracy is somewhat less influenced by the increase in length of the unit since its trajectory was positive with and without controlling for the length of clause while AS unit accuracy had a negative change trajectory before controlling for length of the unit of analysis. Therefore, clause-level accuracy is more straightforward, and researchers are cautioned to control for length of the unit when using any error-free accuracy measure. Further discussion of my evaluation of the accuracy measures is found in Chapter 6, Section 2.3.1.

5.2 COMPLEXITY

Three grammatical complexity measures were calculated in order to capture the development of grammatical complexity. Length of AS unit (C1) was calculated as the mean number of words per AS unit, and clause length (C2) was calculated as the mean number of words per clause. Clauses per AS unit (C3) was calculated as the number of clauses divided by total AS units. Clauses included finite clauses and non-finite clauses with a complement or adjunct. Lexical variety (C4) was measured in D, based on words. In general, I hypothesized growth in all measures. I expected lower initial proficiency to negatively influence initial scores but to increase growth rate. I did not expect differences based on age, gender, or language background or instruction cohort.

5.2.1 Length of AS unit (C1)

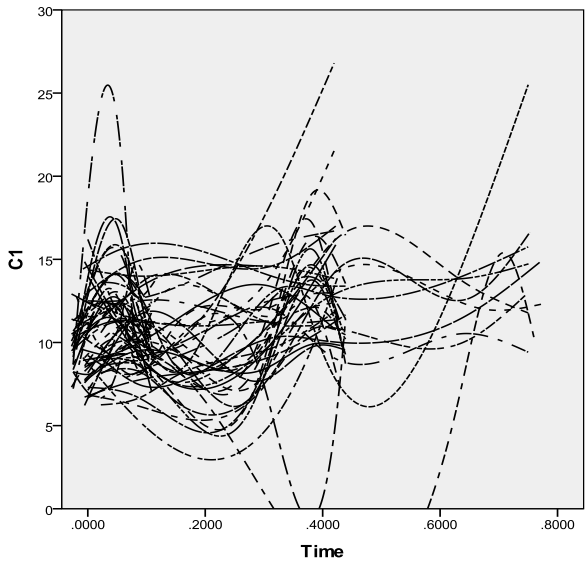

Figure 4 Collection of Smooth Nonparametric and OLS Trajectories of C1 Scores

Figure 4 shows the collection of smooth nonparametric and OLS trajectories across participants (n = 66). Time was adjusted so that the intercept is approximately at the start of data collection. Although Figure 4 suggested the possibility of a non-linear trajectory, a chi-square test was performed and a non-linear trajectory was rejected, ($\chi^2_4 = 3.047$, $p = .550$).

According to the parametric linear model descriptive statistics (Table 20) across these data, the average estimated initial score was 9.97 words per AS unit (SD = 1.26) and the average estimated slope was 6.93 words per AS unit (SD = 4.67) over the year. The magnitude of the standard deviations suggests that the participants differ considerably in their fitted initial scores and in their fitted rates of change. The intraclass correlation for this measure was .1159, which indicates some between-individual clustering.

Table 20 Descriptive Statistics for Individual Growth Parameters of C1 (n=66)

	Initial status (intercept)	Rate of change (slope)
Mean	9.97	6.93
Standard deviation	1.26	4.67

The results of the unconditional linear growth model (Table 21) shows that the expected length of AS unit (C1) at month one was estimated to be 9.974 words for average-aged (25.3 years) and average proficiency (19.2) students with a mean linear growth rate of increase of 6.925 words, indicating a highly significant positive average rate of increase in length of AS unit (C1) over time in the study. Both the mean intercept and growth rate were statistically significant, indicating that both parameters are necessary for describing the mean growth trajectory. Students vary significantly in their length of AS unit (C1) scores at one month in the program ($\chi^2_{65} = 103.095$, $p = .002$) and in growth rate ($\chi^2_{65} = 99.708$, $p = .004$).

Table 21 Unconditioned Linear Model of Growth of Length (in words) of AS unit (C1)

Fixed Effects	**Coefficient**	**SE**	***t***	***p***
Mean Initial status, B_{00}	**9.974**	**.24**	**41.10**	**<.001**
Mean growth rate, B_{10}	**6.925**	**1.04**	**6.65**	**<.001**
Random Effects	**Variance Component**	***df***	X^2	***p***
Initial status, r_{01}	**1.592**	**65**	**103.095**	**.002**
Change rate, r_{1i}	**21.807**	**65**	**99.708**	**.004**
Level-1 error, e_{ti}	**4.865**			

Since there was significant variation in initial scores and in growth rate for this measure, time-invariant predictors were evaluated for inclusion in the model. Results of the best linear growth model include initial proficiency and gender as level-2 predictors for the intercept and gender as a level-2 predictor in the slope model. (Gender was split: 34 males, 32 females.[5]) The

[5] Although genders is not evenly distributed within language groups, a competing model with L1 replacing gender found that L1 was not significant in either the intercept model, $p = .092$ or the slope model, $p = .948$.

expected length of AS unit (C1) of average male students was estimated to be 10.355 words per AS unit (Table 22). On average, female students had lower initial length of AS unit (C1) scores by .954 words. The initial proficiency coefficient suggests that for every unit increase in centered initial proficiency, there was a corresponding increase of initial length of AS unit (C1) scores by .252 words. The mean intercept, gender, and initial proficiency were significantly different from zero, indicating all three parameters are necessary to describe the mean growth trajectory. The mean linear growth rate for male students was 5.276 words/AS unit. Gender was strongly related to growth rate. On average, female students had a higher growth rate by 6.037 words/AS unit.

Table 22 Conditioned Linear Growth Model of Length of AS unit (C1)

Fixed Effects	**Coefficient**	**SE**	***t***	***p***
Model for initial status, π_{0i}				
Intercept, β_{00}	**10.355**	**.292**	**35.507**	**<.001**
Gender, β_{01}	**-.954**	**.401**	**-2.380**	**.020**
Initial Proficiency, β_{02}	**.252**	**.035**	**7.253**	**<.001**
Model for Growth Rate, π_{1i}				
Intercept, β_{10}	**5.276**	**.899**	**5.867**	**<.001**
Gender, β_{11}	**6.037**	**2.174**	**2.777**	**.007**
Random Effects	**Variance Component**	***df***	X^2	***p***
Initial status, r_{01}	**.318**	**63**	**71.895**	**.207**
Change rate, r_{1i}	**7.797**	**64**	**77.380**	**.122**
Level-1 error, e_{ti}	**5.013**			

Students did not vary significantly in their C1 scores at one month (χ^2_{63} = 71.895, p =.207) after controlling for initial proficiency and gender, nor did the students vary significantly in growth rate at one month (χ^2_{64} = 77.380, p = .122) after controlling for gender. By comparing the variance of initial status from Table 21 and Table 22, the proportion of initial score variance explained by gender and initial proficiency combined was calculated to be 80.0%. By comparing the variance of change rate from Table 21 and Table 22, the proportion of change rate variance

explained by gender was calculated to be 64.2%. With no significant variance remaining, no additional analysis (including time-varying predictors) was warranted.

The results indicate that male students had higher initial scores and that within each gender, higher initial proficiency corresponded with higher initial scores. Female students, however, had steeper growth trajectories than male students.

Since the gender difference was unexpected, the data were closely reviewed. As mentioned, few students were enrolled longer than two semesters so few students (only 15.1%) had more than five observations. As such, there were few observations beyond six months after enrollment (n=21) and few observations from females students after six months in the program (n = 5). Therefore, it was possible that these observations do not represent the data as well as the denser earlier observations. Additionally, a female student had a high C1 score (25.5 words) at a later observation. In order to evaluate if the length of AS unit (C1) results, specifically the slope, were disproportionately affected by this potential outlier, I reran the analysis with a limited data set (n=273), excluding observations beyond six months after enrollment. Although gender misses significance (p = .090), the conclusions are the same: female students have lower initial scores but a steeper growth rate. Therefore, the analysis with the full data set was confirmed and thus, retained since it included more observations.

5.2.2 Clause Length (C2)

Figure 5 shows the collection of smooth nonparametric and OLS trajectories across participants of clause length (C2). In fact, the parametric linear model descriptive statistics (Table 23) lists the average mean for clause length (C2), as 5.89 words per clause (SD = .219). This magnitude

of the standard deviation (in comparison to the mean) indicates that the participants are not widely scattered around this initial score.

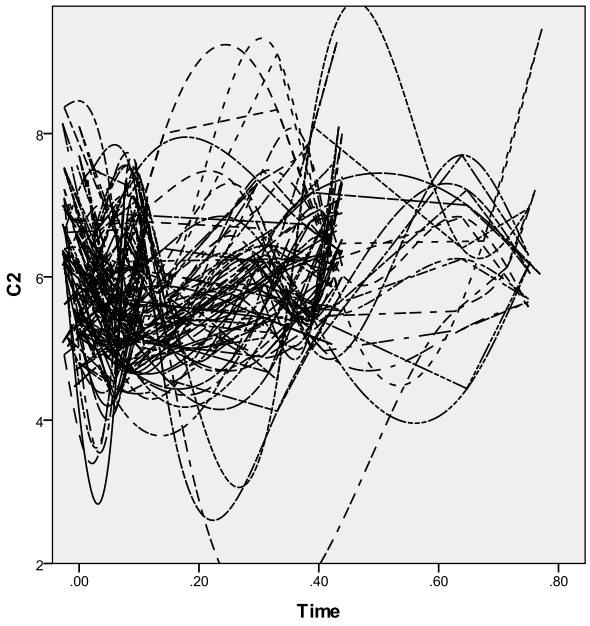

Figure 5 Collection of Smooth Nonparametric and OLS Trajectories of C2 Scores

The average estimated slope was .786 (SD .956). The intraclass correlation for this measure was .0240, which means that the variance explained by between-individual differences (group-level) is only 2.4%. This low intraclass correlation indicates that there was very little clustering. The possibility of a non-linear trajectory was evaluated and rejected, ($\chi^2_4 = 3.543$, $p = .471$).

Table 23 Descriptive Statistics for Individual Growth Parameters of C2 (n=66)

	Initial status (intercept)	Rate of change (slope)
Mean	5.89	.786
Standard deviation	.219	.956

The results of the unconditioned linear growth model (Table 24) states that the expected clause length (C2) was estimated to be 5.892 words for average-aged and average proficiency

students at one month in the program and the mean linear growth rate was estimated to be .786, indicating an average rate of increase in clause length (C2) during the study. Both the mean intercept and slope were statistically significant, indicating that both parameters are necessary for describing the mean growth trajectory. Students did not vary significantly in their C2 scores at one month (χ^2_{65} = 82.379, p = .072) nor in their growth rate (χ^2_{65} = 77.914, p = .131). Since the students did not differ significantly in either initial score or in change rate, further analysis is not warranted. Overall, students increase their mean length of clause during the study. The mean of all students of initial scores is 5.89 words per clause and the expected increase is to 6.68 words per clause after a year.

Table 24 Unconditioned Linear Model of Growth of Mean Clause Length (C2)

Fixed Effects	Coefficient	SE	*t*	*p*
Mean Initial status, B_{00}	5.892	.08	77.265	<.001
Mean growth rate, B_{10}	.786	.30	2.651	.010
Random Effects	Variance Component	*df*	X^2	*p*
Initial status, r_{01}	.048	65	82.379	.072
Change rate, r_{1i}	.914	65	77.914	.131
Level-1 error, e_{ti}	.075			

5.2.3 Clauses per AS unit (C3)

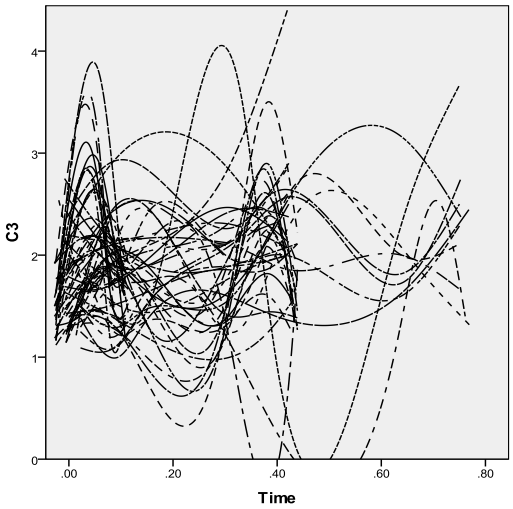

Figure 6 Collection of Smooth Nonparametric and OLS Trajectories of C3 Scores

The collection of smooth nonparametric and OLS trajectories across participants for clauses per AS unit (C3) is illustrated in Figure 6. The parametric linear model descriptive statistics (Table 25) reports that across the data, the average estimated initial score was 1.74 clauses per AS unit (SD = .133) and the average estimated slope was a increase of .779 clauses/AS unit (SD = .568) over the year. The intraclass correlation for this measure was .0522, which means that the variance explained by between-individual differences (group-level) was only 5.2%. This low intraclass correlation indicates that there was very little clustering effect. The possibility of a non-linear trajectory was evaluated and rejected, (χ^{24} = 1.879, p = .758).

Table 25 Descriptive Statistics for Individual Growth Parameters of C3 (n=66)

	Initial status (intercept)	Rate of change (slope)
Mean	1.74	.779
Standard deviation	.133	.568

The results of the unconditioned linear growth model (Table 26) specify that the expected clauses/AS unit (C3) score for an average student was estimated to be 1.737 clauses at one month in the program; and the mean linear growth rate was estimated to be .779, indicating an average rate of increase in clauses/AS unit (C3). Both the mean intercept and slope were statistically different from zero, indicating both parameters are necessary for the mean growth trajectory. Students did not vary significantly in their clauses/AS unit (C3) scores at one month (χ^2_{65} = 69.580, p = .326) nor in their growth rate (χ^2_{65} = 69.005, p = .343). Since the students did not differ significantly in either their initial rate or the change rate, further analysis was not warranted. These results indicated that there is slight growth in this measure (on average from 1.74 clauses per AS unit to 2.52 clauses per AS unit) over the course of the study.

Table 26 Unconditioned Linear Model of Growth in Clause/AS unit (C3)

Fixed Effects	Coefficient	SE	*t*	*p*
Mean Initial status, B_{00}	1.737	.04	41.60	<.001
Mean growth rate, B_{10}	.779	.17	4.49	<.001
Random Effects	Variance Component	*df*	X^2	*p*
Initial status, r_{01}	.018	65	69.580	.326
Change rate, r_{1i}	.322	65	69.005	.343
Level-1 error, e_{ti}	.215			

5.2.4 Lexical Variety (C4)

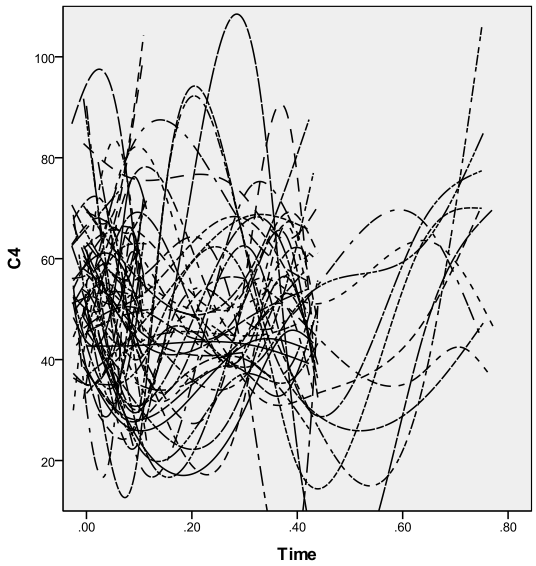

Figure 7 Collection of Smooth Nonparametric and OLS Trajectories of Lexical Variety (C4)

Figure 7 illustrates the collection of smooth nonparametric and OLS trajectories across participants for lexical variety (C4). A non-linear trajectory was confirmed (χ^2_4 = 14.642, p = .006), meaning that the data do not follow a linear trajectory. So, the data were fitted with a quadratic growth model.

Table 27 Descriptive Statistics for Individual Growth Parameters of C4 (n=66)

	Initial status (intercept)	Rate of change (slope)	Rate of Acceleration
Mean	53.72	-26.791	53.89
Standard deviation	7.48	26.789	54.23

The parametric quadratic model descriptive statistics (Table 27) lists that the average estimated initial score across the participants was 53.71 (SD = 7.48) , the average slope was -26.79.62 (SD = 26.79), and the average estimated acceleration was 53.89 (SD = 54.23). The magnitude of the standard deviations suggests that the participants differ considerably in their fitted initial scores,

in their fitted rates of change, and in their fitted rates of acceleration. The intraclass correlation of .198 indicates that there is some between-individual clustering.

Table 28 Unconditioned Quadratic Model of Growth of Lexical Complexity (C4)

Fixed Effects	Coefficient	SE	*t*	*p*
Mean Initial status, β_{00}	53.716	1.429	37.58	<.001
Mean growth rate, β_{10}	-26.791	11.168	-2.40	.019
Mean acceleration, β_{20}	53.891	20.470	2.63	.011
Random Effects	Variance Component	*df*	X^2	*p*
Initial status, r_0	55.913	65	90.202	.021
Change rate, r_1	717.653	65	65.381	.464
Acceleration, r_2	2940.486	65	69.706	.322
Level-1 error, e	168.110			

Results of the unconditional quadratic growth model (Table 28) specify that the expected lexical variety (C4) score at one month for an average student was estimated to be 53.716; the mean linear change rate at one month was estimated to be -26.791, and the mean acceleration of 53.891. All parameters were significantly different from zero, indicating that the three parameters are necessary for describing the mean growth trajectory.

Students varied significantly in their lexical variety (C4) scores at one month (χ^2_{65} = 90.202, p = .021), but did not differ significantly in growth rate (χ^2_{65} = 65.381, p = .464) or acceleration (χ^2_{65} = 69.706, p = .322) at one month (considering the large standard deviations). With significant variation in the initial values for this measure, time-invariant predictors (including instruction cohort, which was previously not predictive) were added to the intercept model in order to explain some of the variation. Since the variance component of the non-linear trajectory was not significant, it was constrained to zero.

Table 29 Conditioned Quadratic Growth Model of Lexical Variety (C4)

Fixed Effects	Coefficient	SE	*t*	*p*
Mean Initial status, β_{00}	43.011	.77	11.40	<.001
Initial Proficiency, β_{01}	.984	.229	4.30	<.001
Cohort, β_{02}	6.186	2.11	2.93	.005
Mean growth rate, β_{10}	-24.730	11.02	-2.25	.026
Mean acceleration, β_{20}	63.337	19.93	3.18	.002
Random Effects	Variance Component	*df*	X^2	*p*
Initial status, r_0	30.168	63	117.399	<.001
Level-1 error, e	173.352			

The results of the conditioned quadratic growth model (Table 29) found initial proficiency and instruction cohort as predictors of initial scores. The expected lexical variety (C4) score for an average student from instruction cohort 1 at one month was estimated to be 43.011. Initial proficiency was strongly related to lexical variety (C4) at one month. The coefficient suggests that for every point increase in (centered) initial proficiency, there was a corresponding .984 increase in lexical variety (C4) scores. Cohort was also strongly related to lexical variety (C4) scores at one month. On average, students in cohort 2 had higher lexical variety (C4) scores by 6.186 points. The mean intercept, initial proficiency, and cohort were significantly different from zero, indicating that that all three parameters were necessary for describing the mean growth trajectory. The mean linear change rate at one month was estimated to be -24.730. The mean acceleration was estimated to be 63.337. Students still varied significantly in their lexical variety (C4) scores at one month (χ^2_{63} = 117.399, p <.001) after controlling for initial proficiency and cohort group. Overall, initial proficiency and cohort explained 46.0% of the variance in initial scores in lexical variety (C4). No other time-invariant predictor was significant.

Since non-linear trajectories are more difficult to visualize than linear growth, Figure 8 shows that the estimated non-linear trajectory for average students from cohort 1 and from cohort

2. Each group showed a slight decrease followed by a steep increase in lexical variety (C4) scores over time. As shown in Table 28, there was not significant variation in the linear change rate or in the acceleration rate, so each group's non-linear trajectory was the same. As described with Table 29, students from instruction cohort 2 had higher lexical variety (C4) scores, than students from cohort 2. Within each cohort, students with higher proficiency had higher lexical variety (C4) scores.

The remaining variation in initial scores might be explained by adding a time-varying predictor. Controlling for topic seems to be appropriate when trying to explain lexical variety. Topic, however, was not useful for describing the mean growth trajectory, and the resulting model did not fit the data as well.

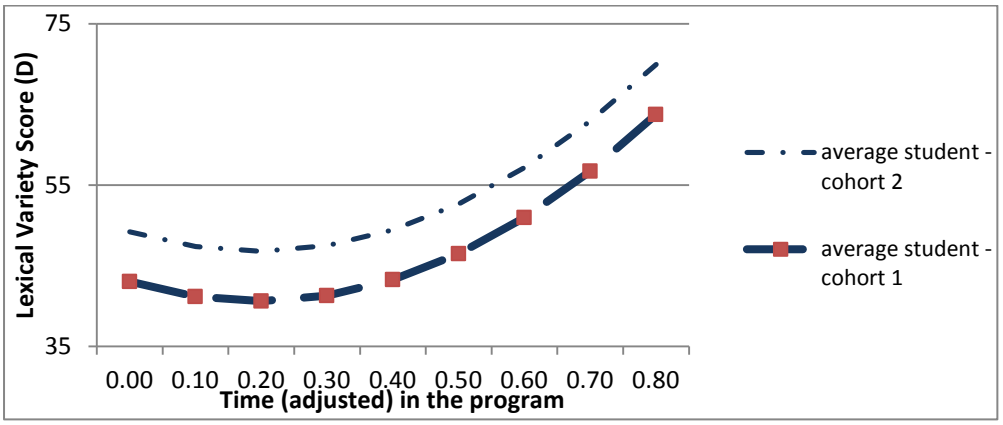

Figure 8 Non-linear Growth Trajectory of Lexical Variety (C4) by Cohort

5.2.5 Correlations among Complexity Measures

I hypothesized a negative correlation between clauses/AS unit (C3) and clause length (C2). I expected a positive correlation between clause length (C2) and lexical variety (C4). I

hypothesized a negative correlation between clauses/AS unit (C3) and lexical variety (C4) and a negative correlation between length of AS unit (C1) and lexical variety (C4).

Table 30 Correlations among the Complexity Measures

	Length of AS unit (C1)	Clause Length (C2)	Clauses/AS unit (C3)	Lexical Variety (C4)
Within-individual Correlations				
Length of AS unit (C1)	--			
Clause Length (C2)	.1221*	--		
Clauses/AS unit (C3)	.8365**	-.4169**	--	
Lexical Variety (C4)	.2014**	-.0533	.1897**	--
Between-individual correlations				
Length of AS unit (C1)	--			
Clause Length (C2)	.2960**	--		
Clauses/AS unit (C3)	.8727**	-.1917	--	
Lexical Variety (C4)	.3096**	.1967	.2093*	--
* $p < .05$ **$p < .01$				

The within-individual and between-individual correlations for the four complexity measures (Table 30) show that length of AS unit (C1) had a strong positive correlation with clauses/AS unit (C3) for both within- (r = .8365) and between-individuals (r = .8727). Obviously, an increase in clauses per AS unit (C3) would increase the overall length of AS unit (C1). Clause length (C2) was also positively correlated with length of AS unit (C1), but the correlations were weaker for the within-individual correlation (r = .1221) and the between-individual correlation (r = .2960). More noteworthy, clause length (C2) and clauses/AS unit (C3) had a moderate negative within-individual correlation (r = -.4169), but the relationship was negligible in the between-individual correlation (r = -.1917). This within-individual correlation finding indicates that an increase in the number of clauses decreases the mean length of clause, but the level-specific negative correlation was not found in the aggregated between-individual correlations.

Length of AS unit (C1) and lexical variety (C4) showed a weak positive correlation both within- (r = .2014) and between-individuals (r =.3096). The positive correlation between length of AS unit (C1) and lexical variety (C4) indicates that students with more lexical variety produced longer sentences. Clause length (C2) and lexical variety (C4) were not correlated, at the within-individual level (r = -.0533) or at the between-individual level (r = .1967). Clauses/AS unit (C3) and lexical variety (C4) were weakly positively correlated, both within-individual (r = .1897) and between individual (r = .2093).

5.2.6 Discussion of Complexity Results

All measures showed growth over time, as hypothesized. Grammatical complexity and lexical complexity, as distinct subcomponents of language complexity, will be discussed separately, followed by a discussion of the correlation data.

5.2.6.1 Grammatical Complexity

The three grammatical complexity measures showed linear growth. An increase in clause length (C2) or in clauses per AS unit (C3) will obviously increase the overall length of the AS (C1). However, with the modest increase in clause length (.79 word or 13.3% of the initial score), the increase in clauses/AS unit (.78 clauses/AS or 44.8% of the initial) has a bigger impact on the increase in AS length (5.28 or 51.0% for males; 6.04 words or 58.3% for females). Length of AS unit (C1) seems to be a practical measure considering the scale of the development over the year.

Initial proficiency impacted initial scores for length of AS unit (C1) (Table 31), as it did with the accuracy measures. However, there was also a difference between males and females in length of AS unit (C1) scores rather than other predictors (such as language background). This

gender difference was unexpected. Although males and females are expected to have equal language learning opportunity in the American program, I failed to anticipate that that males and females might have had different language learning opportunities before entering the program. Although specific information about language learning opportunities is not available, it is plausible that male students generally have greater opportunities to speak in class (Romaine, 2003), perhaps more so in Arabic and East Asian cultures. With this measure, females had a lower initial mean score, but with time in the program, the mean scores became similar to the mean male scores. As such, the result may indicate not a gender difference based on ability, but a gender difference based on a lack of opportunity, which dissipates when language learning opportunities are equal.

Table 31 Summary of Complexity Results

measure	difference in initial score?	difference in slope?	predictors
Length of AS unit (C1)	yes	yes	initial proficiency gender
Clause Length (C2)	no	no	none
Clauses per AS unit (C3)	no	no	none
Lexical Variety (C4)	yes*	no	initial proficiency cohort

*significant variance remains unexplained

There were no differences between students in the clause length (C2) scores and the clauses/AS unit (C3) scores, not even based on initial proficiency, which was contrary to my hypothesis. There was very little variation in clause length and clauses per AS unit scores, perhaps, because of the pressure of second-language online speech production. It might be that ideas are expressed orally in about six words (about the initial score of C2) and clauses are usually limited to a few clauses per AS unit, which makes it easier for the learner to produce. The variance (within-individual and between-individual) in the scores of both of these measures

were low. Nevertheless, the model did capture significant growth even though it was small (less than a word per clause increase in clause length and less than one additional clause/AS unit over a year). Oral production does tend to be shorter than written production. Ahmadian and Tavakoli (2011) had differing conditions aimed at comparing the effects on language complexity, but they also found little variation in clauses per AS unit scores in each of their conditions.

When Norris and Ortega suggested the usefulness of a subclausal measure of complexity, they suggested that this measure is likely to be most predictive for advanced learners as "processes of grammatical metaphor begin to unfold…" (p. 564). It may be that this population has not yet reached an advanced level to show substantial growth in subclausal complexity. The coding of a clause, however, also affected the scores on this measure. Grammatical metaphor processes, such as nominalization, are counted as a separate clause in the Foster et al. (2000) system if the non-finite verb includes a complement (an object) or adjunct. As such, these non-finite verb clauses will increase the number of clauses, rather than increasing the words per clause as suggested by Norris and Ortega. Therefore, this measurement of clause length (as operationalized by Foster, et al.) does not truly capture what Norris and Ortega intended with a subclausal measure. Therefore, a different method of coding clauses might be warranted. One possible coding solution would be to only label finite clauses as clauses, or to code for both, despite the difficulties in consistently separating AS units if all clauses are limited to finite verb clauses as described by Foster et al.

Initial proficiency did not affect the change rate (slope) for any of the grammatical complexity measures, C1, C2, or C3. This finding indicates that all groups improved approximately at the same rate. There was no support for steeper gains for more proficient students, contra Stanovich (1986) and Wendel (2007) in any of the three grammatical complexity

measures. But gender impacted length of AS unit (C1) growth rate, which was not predicted. In general, females produce shorter AS units (9.4 words per AS unit) than males (10.36 words per AS unit) at first but increased their score more (+ 6.037 words) than males after one year in the program, indicating an initial gender difference in this measure was erased with time in the program.

5.2.6.2 Lexical Variety

The lexical variety (C4) results were strikingly different than the grammatical complexity measures, which is not unexpected, since they measure distinct constructs. The results support the inclusion of lexical variety when studying language complexity. The students showed non-linear growth, with a slight decrease and then a steep increase in lexical variety (C4) scores.

Lexical variety (C4) was also influenced differently than the other complexity measures, in that instruction cohort and initial proficiency were relevant predictors. Initial proficiency was an expected predictor, but instruction cohort was an unexpected predictor. It is unclear why instruction cohort would be so highly significant on this measure, specifically, why cohort 2 had statistically higher lexical variety (C4) scores than cohort 1. One possible explanation is that the students in the cohorts differed (particularly with lexical variety performance) upon enrollment, despite uniform placement procedures, or became different early in the semester because of group dynamics.

As shown in Table 8, students from cohort 1 and cohort 2 were not asked to speak on the same topics. Therefore, an alternative explanation for the difference found between cohort 1 and cohort 2 could be a systematic difference in topic effects, which has been found in other research (e.g., Yu, 2009). The lexical variety scores did vary by topic (Table 32), and the scores differed

among topics even within a single semester with the same participants. As such, the topics given to the different cohorts may have influenced the lexical variety scores in unplanned ways.

Table 32 Means (Standard Deviation) of Lexical Variety (C4) Scores by Topic

cohort 1			cohort 2	
low-intermediate	high-intermediate	advanced	low-intermediate	high-intermediate
childhood meal M = 46.86 SD =15.22	world problem M = 54.83 SD =17.47	media violence M = 53.39 SD =12.41	best friend M = 53.81 SD =12.97	ideal vacation M = 46.43 SD = n.a
transportation M = 30.73 SD = 7.43	a regret M = 41.13 SD =13.23	computerized society M = 60.11 SD =8.61	a surprise M = 52.56 SD =17.72	renting M = 34.21 SD = n.a
admired person M = 45.52 SD =12.06		internet risks M = 65.33 SD =19.01		home city M = 50.63 SD =13.06
		extravagant lifestyle M = 51.52 SD =9.01		ideal job M = 50.79 SD =11.20
		rich and poor M = 59.44 SD =19.53		disliked custom M = 58.10 SD = 12.92
				famous person M = 54.61 SD =16.16

As discussed, the lexical variety scores did not show a linear pattern in growth, but we might expect the mean lexical variety scores from the topics from the low-intermediate level to be lower than the mean scores at the high-intermediate level which are lower than the mean scores from the advanced level. That pattern is found generally in the means from cohort 1, in that scores from the high-intermediate level are generally higher than the low-intermediate scores but lower than the scores from the advanced level. For cohort 2, however, the lexical variety scores from at the low-intermediate level, are a rather high (ranking as the twelfth and tenth highest overall from the eighteen topics) and the scores at the high-intermediate level do not increase much, if at all.

Even more importantly, the standard deviations given in Table 32 shows how the lexical variety (C4) scores of some topics were more tightly clustered (as shown by the low standard

deviations) than for other topics (with much higher stand deviations). Thus, the strength of topic effect varied. For instance, the topic discussing different modes of transportation had the lowest lexical variety (C4) median score and the scores were tightly clustered. Other topics show a larger range of scores, where (considering the results of the best fitting model in Table 29) more proficient students may have much higher scores than other students. Importantly, the impact of the topic effects was inconsistent, which means that controlling for topic was less successful than controlling for a constant effect (e.g., initial proficiency). Moreover, the number of speeches differs per topic (as listed in Table 8) which again makes assessing topic effects tenuous.

Regardless, in an attempt to determine if topic effects explained the difference in lexical variety (C4) scores, two additional models were evaluated: a model with topic added to the best-fitting model and a model with topic instead of cohort. When topic was added to the model with cohort, topic was not significant predictor (p=.147), but more of the variance in the data was explained, which means the model was improved. However, the comparison between the simpler model (with only Level-2 predictors) and the more complex model (with Level-2 predictors and the time-varying predictor of topic) was not significant, which means that topic as a predictor does not explain enough of the variance to off-set the increase in the number of parameters.

It is possible that adding topic to the model did not reach significance because instruction cohort already explained some of the topic effect variation. Therefore, cohort was deleted from the model and the time-varying predictor topic was added. Again, topic was not a significant predictor (p=.885), and the resulting model explained less of the variance in the data. This indicates that the cohorts did simply differ in lexical variety performance; perhaps cohort 2 had a greater command of English vocabulary in general. Overall, the difference in lexical variety (C4) scores is likely driven by some inconsistent topic effects, or that the topics given to the cohorts

elicited greater variety of lexical items, and there was a difference in the populations of cohort 1 and cohort 2, which was larger than a difference in initial proficiency. Regardless, all students improved over the course of the study. These findings call for a piloting of topics to further investigate if likely inconsistent topic effects can be controlled, especially if comparisons across instruction cohorts are important.

5.2.6.3 Correlations within Complexity

No hypotheses were made about the correlations between length of AS unit (C1) and clause length (C2) or about the correlation between length of AS unit (C1) and clauses per AS unit (C3) because mathematically an increase in either of the sub-measures would also increase the C1 score. My hypothesis was that there would be a negative correlation between clause length (C2) and clauses per AS unit (C3), and the results showed a negative correlation between these measures. However, based on the overall picture, the reasoning behind the hypothesis was incorrect. The negative correlation was not driven by the students' increase in nominalization and other more metaphoric language because with the coding system, non-finite clauses (such as nominalization) are counted as a clause. Since non-finite clauses in the coding system can be as little as three words, an increase in these types of clauses pulled down the mean clause length, especially seen in the within-individual correlations.

I hypothesized that length of AS unit (C1) and lexical variety (C4) would be negatively correlated because a varied vocabulary might be so challenging for lower proficiency students that they would produce short AS units to off-set the difficulty in lexical items (based on the mixed results found by Skehan, 2009a). The hypothesis made concerning lexical variety (C4) and length of AS unit were not supported. Length of AS unit (C1) and lexical variety (C4) scores had a weak to moderate *positive* correlation. The stronger between-individual C1-C4 correlation

suggests that for some students, they are connected growers, but not for all students. This might be because overall proficiency drives both measures.

Clause length (C2) and lexical variety (C4) did not show a correlation; clause length (C3) and lexical variety (C4) had a weak, but positive correlation. The positive correlation between lexical variety (C4) and both length of AS unit (C1) and with clauses/AS unit (C3) might be another function of underlying proficiency. The lack of correlation between lexical variety (C4) and clause length (C2) is strange because one would think that lexical choices would affect the length (either negatively or positively) at the clause level. This finding seems to indicate that the impact of lexical choices is not at the clause level, as measured in this study.

The results support Skehan's claim that lexical diversity should be included in descriptions of language complexity. Minimally, the results showed that lexical diversity follows a different path of development than the linear growth of the grammatical complexity measures and it is more affected by topic effects than the other measures.

There was also support for measuring complexity by length of AS unit and clause length. These two measures were not highly correlated which indicates that both are useful in understanding development. There were high correlations between these measures and clauses per AS unit (C3), which indicates this measure was less useful when the other two are already calculated. When coding clauses by the Foster et al. (2000) suggestions, it is recommended that the complexity by subordination be calculated by number of finite clauses per AS units. This recommendation is made so that the subordination measure is more in line with the theoretical underpinnings, specifically, an increase in complex language such as nominalizations would increase mean clause length and rather than be treated as a short non-finite clauses which artificially lowers mean length of clause. In addition, this recommended change would allow

researchers to study the development of finite and non-finite clauses separately, since the subordination ratio would be based only on finite clauses.

5.3 FLUENCY

There were three measures of fluency: phonation time ratio (F1) (a general measure of fluency), mean length of pause (F2) (a measure of fluency breakdown), and mean length of fluent run (F3) (a measure of fluency proceduralization). Mean length of pause was calculated as the average pause length of filled (e.g., “uh”) and unfilled (i.e., silent) pauses of 200 milliseconds or more. Mean length of fluent run was calculated as the average stretch of speech, in syllables, bounded by pauses of at least 200 milliseconds. An increase in phonation time ratio (F1) or in mean length of fluent run indicates improving fluency, as does a decrease in mean length of pause (F2). I hypothesized growth in all measures, with initial proficiency leading to higher initial scores but flatter growth trajectories.

5.3.1 Phonation Time Ratio (F1)

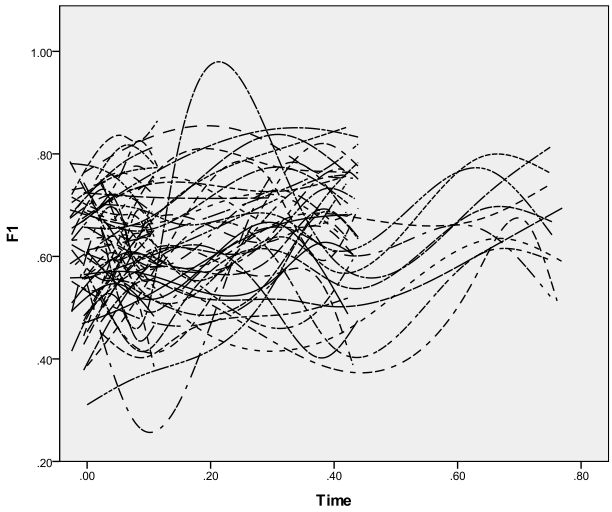

Figure 9 Collection of Smooth Nonparametric and OLS Trajectories of F1 Scores

Figure 9 shows the collection of smooth nonparametric and OLS trajectories across participants (n = 66) of phonation time ratio (F1) with time adjusted so that the intercept is approximately at the start of data collection. Although the plot (Figure 9) indicated the possibility of a nonlinear change trajectories, the quadratic model was rejected as not significantly different from the linear model ($\chi^2_4 = 7.340$, $p = .119$).

Table 33 Descriptive Statistics for Individual Growth Parameters of F1 (n=66)

	Initial status (intercept)	Rate of change (slope)
Mean	.601	.209
Standard deviation	.093	.200

The parametric linear model descriptive statistics (Table 33), with the average estimated initial scores was .601 (SD = .093) and the average estimated slope of .209 (SD = .200), indicates that the average student in this population was able to increase his phonation time ratio

(speaking time) by .209. The intraclass correlation was .4172, indicating the data had between-individual clustering.

Table 34 Unconditioned Linear Model of Growth of Phonation Time Ratio (F1)

Fixed Effects	Coefficient	SE	*t*	*p*
Mean Initial status, B_{00}	.601	.01	47.23	<.001
Mean growth rate, B_{10}	.209	.04	5.60	<.001
Random Effects	Variance Component	*df*	X^2	*p*
Initial status, r_{01}	.009	65	277.883	<.001
Change rate, r_{1i}	.040	65	141.876	<.001
Level-1 error, e_{ti}	.004			

Results of the unconditional growth model (Table 34) specify that the expected phonation time ratio (F1) score for an average-aged (25.3 years), average initial proficiency (19.2 on the placement test) student was estimated to be .601 at one month in the program, and the mean linear growth rate was estimated to be .209. Both the mean intercept and slope were statistically significant, indicating that both parameters were necessary for describing the mean growth trajectory model. Students varied significantly in their phonation time ratio (F1) scores at one month ($\chi^2_{65} = 277.883$, $p < .001$) and in their growth rate ($\chi^2_{65} = 141.876$, $p < .001$).

Since there was significant variation in the initial values and in the change trajectory for this measure, time-invariant predictors were added to the model. The best fitting model had initial proficiency in the intercept model only. No other time-invariant predictor (i.e., age, gender, L1, cohort) was significant.

The results with predictors (Table 35) list that the expected phonation time ratio (F1) for an average student at one month was estimated to be .600. Initial proficiency was strongly related to phonation time ratio (F1) at one month. For every point increase in centered initial proficiency, the phonation time ratio (F1) scores are expected to increase by .008. The mean linear growth rate at one month was estimated to be .226, indicating a positive average rate of

increase in phonation time ratio (F1) (i.e., the student produced more language) during the study. Students varied significantly in their F1 scores at one month (χ^2_{64} = 235.622, p < .001) and in growth rate at one month (χ^2_{65} = 143.337, p < .001) after controlling for initial proficiency in the intercept. By comparing the variance components in Table 34 and Table 35, the proportion of intercept variance explained by initial proficiency was calculated to be 22.2%.

Table 35 Conditioned Linear Growth Model of Phonation Time Ratio (F1)

Fixed Effects	Coefficient	SE	*t*	*p*
Model for initial status, π_0				
Intercept, β_{00}	.600	.012	50.97	<.001
Initial Proficiency, β_{01}	.008	.002	3.99	<.001
Model for Growth Rate, π_{1i}				
Intercept, β_{10}	.226	.038	5.95	<.001
Random Effects	Variance Component	*df*	X^2	*p*
Initial status, r_{01}	.007	64	235.622	<.001
Change rate, r_{1i}	.039	65	143.337	<.001
Level-1 error, e_{ti}	.004			

Time-varying predictors should only be added to the model if there is a clear directional prediction. As such, the accuracy measures and grammatical complexity measures are not valid variables to control because it is unclear how high or low accuracy would affect phonation time ratio or how long or short utterances would affect phonation time ratio. Both accuracy measures, however, were added to the model to check test if accuracy was a predictive variable in fluency, and were found not to be significant predictors. Each of the complexity measures were also added in turn. Both clauses/AS unit (C3) and lexical variety (C4) improved the model slightly with little difference between them. In both of these potential models, a higher complexity score predicted a higher F1 score. Lexical variety (C4) might be valid predictor to control while looking at phonation time ratio.

Table 36 Conditioned Linear Growth Model of Phonation Time Ratio (F1) with C4

Fixed Effects	Coefficient	SE	*t*	*p*
Model for initial status, π_{0i}				
Intercept, β_{00}	.559	.025	22.506	<.001
Initial Proficiency, β_{01}	.008	.002	3.60	.001
Model for Growth Rate, π_{1i}				
Intercept, β_{10}	.223	.038	5.83	<.001
Model for C4 Growth Rate, π_{21i}	.001	<.001	2.00	.050
Random Effects	Variance Component	*df*	X^2	*p*
Initial status, r_{01}	.017	35	85.918	<.001
Change rate, r_{1i}	.044	36	124.436	<.001
C4 change rate, r_{2i}	<.001	36	56.790	.015
Level-1 error, e_{ti}	.004			

The model with time-varying lexical variety (C4) (Table 36) was found to be significantly better than the conditioned model without any time-varying predictors (χ^2_7 = 15.558, p < .03). For average students with average lexical variety (C4) scores, the expected phonation time ratio (F1) score was estimated to be .559. Proficiency was strongly related to phonation time ratio (F1) scores at one month. For every increase in one point in centered initial proficiency, there was a corresponding increase of phonation time ratio (F1) scores by .008. The mean linear growth rate at one month was estimated to be .223, while controlling for lexical variety (C4). At a certain time point, one unit increase in lexical variety (C4) scores increased the phonation time ratio (F1) scores by .001. All were significantly different from zero, indicating that all parameters are necessary for describing the mean growth trajectory. Students varied significantly in their F1 scores at one month (χ^2_{35} = 85.918, p < .001), in growth rate at one month (χ^2_{36} = 124.436, p < .001), and in the relationship between phonation time ratio and lexical variety (C4) scores (χ^2_{36} = 56.790, p = .015), after controlling for proficiency in the intercept and lexical variety (C4) scores.

Generally, higher proficiency students are able to spend more time speaking based on F1 scores than lower proficiency students. Interestingly, when students have higher lexical variety

(C4) scores, they have even higher overall speaking time (F1) scores. This finding means that the use of varied lexical items does not decrease fluency; an increase in lexical variety corresponds with an increase in fluency.

5.3.2 Mean Length of Pause (F2)

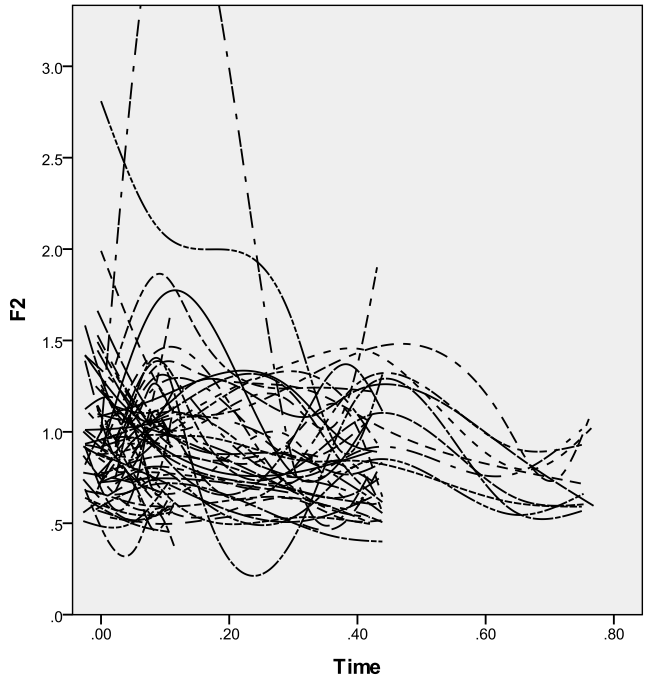

Figure 10 Collection of Smooth Nonparametric and OLS Trajectories of F2 Scores

The collection of smooth nonparametric and OLS trajectories across participants for mean length of pause (F2) is in Figure 10. Time was adjusted so that the intercept is approximately at the start of data collection. Although the plot indicated the possibility of nonlinear trajectories, a non-linear model was rejected, ($\chi^2_4 = 3.056$, $p = .549$).

The parametric linear models descriptive statistics (Table 37), with the average estimated initial score of .997 (SD =.316) and the average estimated slope of -.485 (SD = .518), indicates

that the average student in the study has an observed mean length of pause of just under a second and decreases his mean length of pause (improved fluency) by .485 over the year. The magnitude of the standard deviations suggests that the participants differ considerably in their fitted initial scores and in their fitted rates of change. The intraclass correlation for this measure was .325, meaning the data showed between-individual clustering.

Table 37 Descriptive Statistics for Individual Growth Parameters of F2 (n=66)

	Initial status (intercept)	Rate of change (slope)
Mean	.997	-.485
Standard deviation	.316	.518

Results of the unconditional growth model (Table 38) shows that both the mean intercept (.997) and slope (-485) for an average-aged, average proficiency student were statistically significant, indicating that both parameters are necessary to describe the change trajectory. Students varied significantly in their mean length of pause (F2) at one month ($\chi^2_{65} = 246.596$, $p < .001$) and in their growth rate ($\chi^2_{65} = 97.827$, $p = .005$).

Table 38 Unconditioned Linear Model of Growth of Mean Length of Pause (F2)

Fixed Effects	Coefficient	SE	*t*	*p*
Mean Initial status, B_{00}	.997	.04	22.553	<.001
Mean growth rate, B_{10}	-.485	.11	-4.639	<.001
Random Effects	Variance Component	*df*	X^2	*p*
Initial status, r_{01}	.100	65	246.596	<.001
Change rate, r_{1i}	.268	65	97.827	.005
Level-1 error, e_{ti}	.064			

Since there was significant variation in the initial values and for change rate for this measure, time-invariant predictors were added to the model to explain the variance. Results of the best-fitting linear growth model with initial proficiency as a level-2 predictor in the intercept only (Table 39) specify that the expected mean length of pause (F2) of an average-aged average

proficiency student at one month was estimated to be 1.005 seconds. Initial proficiency was strongly related to F2 scores at one month. On average, one point increase in centered initial proficiency score corresponded with a shorter mean length of pause (F2) by .022 seconds. The mean linear growth rate of all students was estimated to be -.594 seconds, indicating a trend for average rate of decrease in mean length of pause (improvement) during the study. Students still varied significantly in their mean length of pause at one month ($\chi^2_{64} = 221.512$, $p < .001$), and in growth rate ($\chi^2_{65} = 96.470$, $p = .005$) after controlling for initial proficiency in the intercept.

Table 39 Conditioned Linear Model of Growth of Mean Length of Pause (F2)

Fixed Effects	Coefficient	SE	*t*	*p*
Model for initial status				
Intercept, β_{00}	1.005	. 042	23.92	<.001
Initial Proficiency, β_{01}	-.022	.007	- 3.03	.004
Model for growth rate				
Intercept, β_{10}	-.594	.121	-4.928	<.001
Random Effects	Variance Component	*df*	X^2	*p*
Initial status, r_{01}	.084	64	221.512	<.001
Change rate, r_{1i}	.267	65	96.470	.007
Level-1 error, e_{ti}	.063			

The model was expanded to include time-varying predictors because significant variance in both initial scores and change rate remained after adding the only relevant time-invariant predictor (based on testing all variables) of initial proficiency. The best fitting model included adding the covariate lexical variety (C4) to the mean length of pause (F2) model. This time-varying predictor is a valid variable to control because it is plausible that students who try to vary their vocabulary choices might have longer pauses.

Results of the linear growth model with time-varying covariate of lexical variety (C4) (Table 40) include initial proficiency and lexical variety (C4) scores as predictors. For average students with average lexical variety (C4) scores at one month in the program, the expected mean

length of pause (F2) score was estimated to be 1.163 seconds. The coefficients suggest that for every increase in one point of centered initial proficiency, there was a corresponding decrease (improvement) in mean length of pause scores by .017 seconds. The mean linear growth rate was estimated to be -.609 second (a decrease in mean pause length) while controlling for lexical variety (C4). At one certain time point, one point increase in lexical variety (C4) score further decreased mean length of pause (F2) by .003 seconds. All are significantly different from zero, indicating that all parameters are necessary for describing the mean growth trajectory. Students varied significantly in their mean length of pause (F2) scores at one month (χ^2_{35}= 286.097, p < .001), change rate (χ^2_{36} = 152.862, p < .001), and in the relationship between mean length of pause and lexical variety (C4) at a given time point (χ_{36} = 188.769, p = .001) while controlling for lexical variety (C4) scores.

Table 40 Conditioned Linear Model of Growth in Mean Length of Pause (F2) with C4

Fixed Effects	Coefficient	SE	*t*	*p*
Model for initial status				
Mean F2 at 1 month, β_{00}	1.163	. 104	11.231	<.001
Initial Proficiency, β_{01}	-.017	.007	-2.28	.026
Mean growth rate, β_{10}	-.609	.125	-4.87	<.001
Mean C4 growth rate, β_{20}	-.003	.002	-2.01	.048
Random Effects	Variance Component	*df*	X^2	*p*
One-month status, r_{01}	.473	35	286.097	<.001
Change rate, r_{1i}	.506	36	152.862	<.001
C4 slope	.0008	36	188.769	<.001
Level-1 error, e_{ti}	.030			

In sum, initial proficiency and lexical variety (C4) scores predicted mean length of pause (F2) scores. Higher initial proficiency corresponded with slightly shorter pauses (smaller mean length of pause scores), regardless of lexical variety (C4) scores. Higher lexical variety (C4) scores corresponded with slightly shorter pauses (lower mean length of pause scores), indicating that retrieval of varied lexical items did not require longer pauses, as might be expected. As with

phonation time ratio (F1), an increase in lexical variety correlated with an improvement in fluency.

5.3.3 Mean Length of Fluent Run (F3)

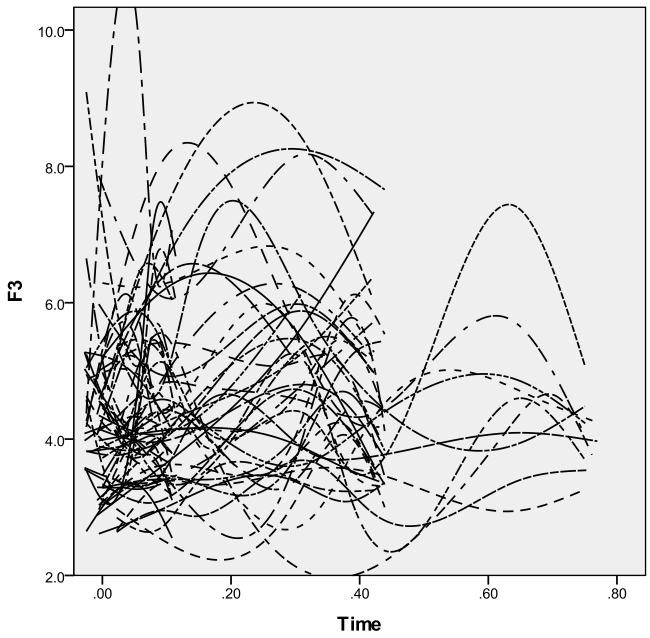

Figure 11 Collection of Smooth Nonparametric and OLS Trajectories F3 Scores

The collection of smooth nonparametric and OLS trajectories across participants for mean length of fluent run (F3) is in Figure 11. The unconditional linear growth model was compared to the quadratic growth model, which found the quadratic growth model to be a better fit for the data (χ^2_4 = 15.147, p =.004). The intraclass correlation for this measure was .4172, which suggests there was between-individual clustering.

Table 41 Descriptive Statistics for Individual Growth Parameters of F3 (n=66)

	Initial status (intercept)	Rate of change (slope)	Rate of Acceleration
Mean	4.31	2.31	-3.32
Standard deviation	.919	6.24	8.32

The parametric quadratic model (Table 41), lists the average estimated initial scores of 4.31 (SD = .919) syllables per fluent run (bounded by pauses of 200ms or more), the average estimated change rate of 2.31 (SD = 6.24), and the average estimated acceleration rate of -3.32 (SD 8.32). The magnitude of standard deviation in change rate and in rate of acceleration (in comparison to the means) suggests that the trajectories are widely scattered, as seen in Figure 11.

Table 42 Unconditioned Quadratic Model of Growth in Mean Length of Fluent Run (F3)

Fixed Effects	Coefficient	SE	*t*	*p*
Mean Initial status, β_{00}	4.309	.133	32.37	<.001
Mean growth rate, β_{10}	2.311	1.037	2.23	.029
Mean acceleration, β_{20}	-3.317	1.482	-2.24	.029
Random Effects	Variance Component	*df*	X^2	*p*
Initial status, r_{01}	.844	65	190.867	<.001
Change rate, r_{1i}	38.884	65	89.920	.022
Acceleration, r_{2i}	69.181	65	69.803	.319
Level-1 error, e_{ti}	.512			

Results of the unconditional quadratic growth model (Table 42) show that parameters (initial score, growth rate, and acceleration) were significantly different from zero, indicating that all parameters are necessary for describing the mean growth trajectory. Students varied significantly in their mean length of fluent run (F3) scores at one month (χ^2_{65} = 190.867, p < .001), and in growth rate at one month (χ^2_{65} = 89.920, p =.020), but they did not differ significantly in acceleration (χ^2_{65} = 69.803, p =.319).

With much of the variance still unexplained, time-invariant predictors were added to the model. The best fitting model included initial proficiency, language background, and age, but each of the additional variables was only predictive for the initial scores.

Results of the quadratic growth model with initial proficiency, L1, and age as level-2 predictors in the intercept model (Table 43) specify that the expected mean length of fluent run (F3) for average-aged, average proficiency Arabic L1 students at one month was estimated to be

4.521 syllables. Initial proficiency, L1, and age were strongly related to mean length of fluent run (F3) at one month. On average, non-Arabic students had lower initial F3 scores by .626 syllable. For every increase in centered age, there was a corresponding decrease in initial mean length of fluent run (F3) score by .053 syllable. And, for every one unit increase in centered initial proficiency score, there was a corresponding increase in initial mean length of fluent run (F3) scores by .100 syllable. The mean linear growth of F3 was 2.117 with a mean acceleration of -2.436. All parameters were significantly different from zero, indicating that all parameters are necessary for describing the mean growth trajectory.

Table 43 Conditioned Non-linear Growth Model of Mean Length of Fluent Run (F3)

Fixed Effects	Coefficient	SE	*t*	*p*
Model for initial status, π_{0i}				
Intercept, β_{00}	4.521	.138	32.69	<.001
L1, β_{01}	-.626	.199	-3.15	.002
Age, β_{02}	-.053	.020	-2.61	.011
Initial Proficiency, β_{03}	.100	.021	4.86	<.001
Mean growth rate, β_{10}	2.117	.955	2.22	.030
Mean acceleration, β_{20}	-2.436	1.240	-1.96	.054
Random Effects	Variance Component	*df*	X^2	*p*
Initial status, r_{01}	.465	62	123.898	<.001
Change rate, r_{1i}	28.631	65	88.646	.027
Acceleration rate, r_{2i}	37.246	65	67.753	.383
Level-1 error, e_{ti}	.524			

By comparing the variance components in Table 42 and Table 43, the proportion of variance in initial score explained by the time-invariant predictors (initial proficiency, L1 and age) was estimated to be 44.9%. Yet, students still varied significantly in their mean length of fluent run (F3) scores at one month (χ^2_{62} = 123.898, p < .001) after controlling for the given level-2 predictors and in growth rate at one month (χ^2_{65} = 88.646, p = .027).

Figure 12 shows the non-linear change trajectories, as described in Table 43. All students had the same trajectory, with rising and then leveling mean length of fluent run (F3) scores with

time in the program. Students with an Arabic language background had higher scores than students with a non-Arabic language background. Within each language background group, higher proficiency students had higher scores than lower proficiency students. Within the proficiency levels, younger students had higher mean length of fluent run (F3) scores than older students.

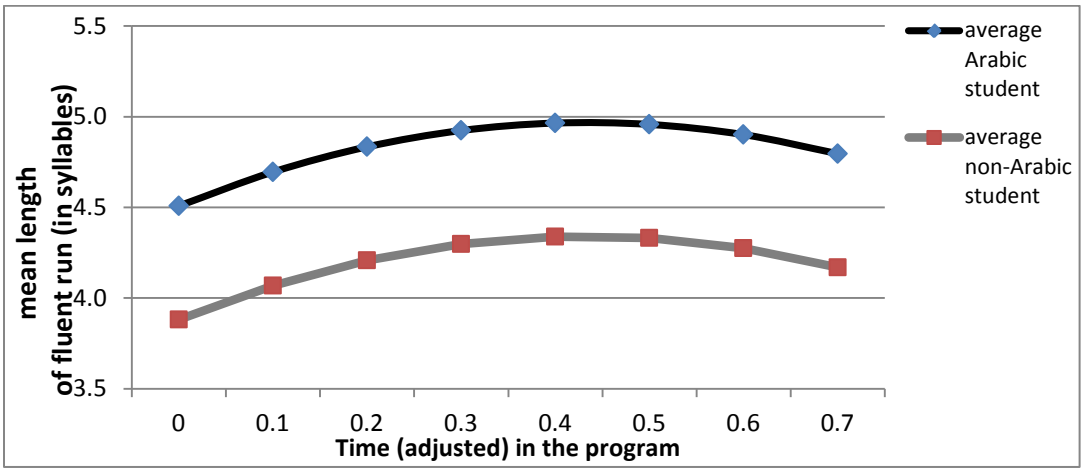

Figure 12 Non-linear Growth Trajectory of F3 of Average Students by L1

5.3.4 Correlations among Fluency Measures

The within-individual and between-individual correlations for the fluency measures (Table 44) show that, unsurprisingly, phonation time ratio (F1) had a strong negative correlation with mean length of pause (F2), which means that when the pause length decreases (improved fluency), the phonation time ratio increases (improved fluency). The within-individual correlation (r = -.8307) and the between-individual correlation (r = -.8301) were significant. Phonation time ratio (F1) scores had a moderate positive correlation with mean length of fluent run (F2), with a within-individual correlation (r = .4919) and a between-individual correlation (r = .5014). An increase

in mean length of fluent run (number of syllables bounded by pauses of 200ms) was correlated with an increase in overall phonation time ratio (F1). These relationships were fully expected, considering that phonation time ratio includes the information about pausing and speech length.

Table 44 Correlations among Fluency Measures

	PTR (F1)	MLP (F2)	MLFR (F3)
Within-individual Correlations			
Phonation Time Ratio (F1)	--		
Mean Length of Pause (F2)	-.8307**	--	
Mean Length of Fluent Run (F3)	.4919**	-.2409**	
Between-individual Correlations			
Phonation Time Ratio (F1)	--		
Mean Length of Pause (F2)	-.8401**	--	
Mean Length of Fluent Run (F3)	.5014**	-.3227**	--
$^{**}p < .01$			

Mean length of pause (F2) and mean length of fluent run (F3) had a weak to moderate negative correlations within-individual (r = -.2409) and between-individual (r = -.3227). This negative correlation means that as the stretch of speech increases, the mean length of pause decreases (both improve fluency), or when mean length of pause (F2) increases, the mean length of fluent run (F3) decreases (both decrease fluency). To be clear, these measures are calculated separately. As such, an increase in one could be correlated with an increase or a decrease in the other. These data showed that an increase in fluency of one of these measures was correlated with an increase in fluency in the other measure. In other words, improvement in fluency proceduralization (fluent runs) did not come at the expense of fluency breakdown (pausing).

5.3.5 Discussion of Fluency Results

5.3.5.1 Predictors of Fluency

As hypothesized, initial proficiency did predict initial scores on all three fluency measures (Table 45). As would be expected, higher initial proficiency correlated with better fluency scores. More interestingly, no other time-invariant in the study predictor was significant for phonation time ratio (F1) or mean length of pause (F2), despite significant variation in student scores. Since variation remained after testing all available time-invariant and time-varying predictors, the findings indicate that (all three of) the fluency measures were influenced by variable(s) not included in the study. Possible variables might be interest in the topic, motivation, or the specific subsection of extraversion of assertiveness.

Table 45 Summary of Fluency Results

measure	difference in initial score?	difference in slope?	predictors
Phonation Time Ratio (F1)	yes*	no*	initial proficiency lexical variety (C4)
Mean Length of Pause (F2)	yes*	no*	initial proficiency lexical variety (C4)
Mean Length of Fluent Run (F3)	yes*	no*	initial proficiency language background age

*significant variance remains unexplained

Phonation time ratio (F1) and mean length of pause (F2) were both influenced by lexical variety (C4) scores. Contrary to expectations, when controlling for lexical variety (C4), students with *higher* lexical variety (C4) scores had higher phonation time ratio (F1) scores, which indicates that a producing a varied vocabulary did not decrease fluency. Moreover, this finding was true for all students. Likewise, lexical variety (C4) was found to be predictive of mean

length of pause (F2). The data showed that when students produced speeches with higher lexical variety (C4) scores, they actually had lower mean length of pause (F2) scores (*shorter* pauses), regardless of initial proficiency. Again, the impact of using a more varied vocabulary was not in the expected direction. In sum, the phonation time ratio model (F1) and the mean length of pause (F2) model were both improved by including lexical variety (C4) scores and the direction was that higher lexical variety (C4) increased the fluency scores. Therefore, these findings did not support a trade-off effect where choosing to use a more varied vocabulary comes at the expense of fluency. High lexical variety (C4) scores neither caused an increase in pauses nor an overall decrease in phonation time ratio. It might be that adding lexical variety (C4), as measured by D, to the model strengthens the model because it serves as another predictor of proficiency. In addition, since lexical variety was not predictive in the model for mean length of fluent run, these findings may indicate that lexical retrieval occurs before the syntactic frames for the utterances are created, supporting MacWhinney's competition model (2001)

Initial mean length of fluent run (F3) scores was influenced by initial proficiency as well as the time-invariant predictors of age and language background. Again, all students had a similar change trajectory, but there were differences in initial scores among different populations within the program. Arabic students had higher mean length of fluent run (F3) scores than non-Arabic students. Age was predictive of length of fluent run (F3) scores. This finding was especially surprising because the age range was not large; all participants were young adults 18 - 35 years.

5.3.5.2 Unexplained Variance

Further, the model showed that there was still significant unexplained variance in both initial score and, importantly, in change rate for each of the fluency measures. Although this

dissertation did not include any affective measures, taken together, this combination of predictors might all reflect a difference in assertiveness that has been connected to increased oral ability (Ockey, 2011). It is plausible that younger adult students might be more assertive than slightly older L2 learners; more proficient students are likely to be more assertive than lower proficiency students; and student with an Arabic language background might be more assertive than non-Arabic students (Chinese and Korean students), considering the cultural background (rather than language background). Regardless, this possible explanation cannot be verified since no affective measures were taken of the students. Of course, other explanations are also possible (e.g., there might be a slight processing advantage with a decrease in age).

5.3.5.3 Growth Trajectories

Once in the program, all students had a similar growth trajectory. Phonation time ratio (F1) and mean length of pause (F2) both had linear growth trajectories whereas mean length of fluent run (F3) had a non-linear trajectory rather than a linear growth trajectory. As noted, mean length of fluent run (F3) is the average number of syllables spoken in a fluent run bounded by pauses of at least 200 milliseconds. The model showed that the mean length of fluent run (F3) scores, peak around four and a half to five syllables and the slope levels to around four syllables. It is important to understand that the model indicates that the slope's growth rate slows down over time, not that this fluency measure actually decreased. It might be that the mean length of an utterance for second language learners does not continue to rise over five syllables before a short (even native-like pause) of 200 milliseconds. Therefore, the appropriate conclusion is that the acceleration of growth in mean length of fluent run (F3) flattens.

In order to look at the data by instruction level, the data were separated by level (low-intermediate, high-intermediate, advanced) and observation[6], and the average mean length of fluent run (F3) score was calculated. In fact, the ending group mean length of fluent run for low-intermediate students was 4.26 syllables, the ending group mean was 4.32 for high-intermediate students, and then to 5.05 for advanced students. So, as a group, the students continue to produce longer stretches of speech level by level.

5.3.5.4 Correlations within Fluency

The direction of the correlations with the overall fluency measure, phonation time ratio (F1), and the other fluency measures was wholly expected. Interestingly and importantly, mean length of pause (F2) and mean length of fluent run (F3) had a negative correlation. Since a decrease in mean length of pause (F2) and an increase in mean length of fluent run (F3) both indicate improving fluency, these data did not support a trade-off effect between mean length of pause (F2) and mean length of fluent run (F3). That is, these data did not support the idea that students take longer pauses in order to produce longer fluent runs. Rather shorter pauses were correlated with longer fluent runs. Students improved their fluency in both ways during development.

5.4 CORRELATIONS BETWEEN CAF CONSTRUCTS

My third research question was concerned with the relationships between CAF constructs. In this section, the correlations between constructs are given.

[6] The full group means data are found in Appendix D.

5.4.1 Accuracy and Complexity

I hypothesized a negative relationship would be found between accuracy and grammatical complexity, following Skehan's theory of tension between control (accuracy) and risk-taking (complexity). In contrast, I expected a positive correlation between lexical variety and accuracy.

5.4.1.1 Accuracy and Grammatical Complexity

The within-individual and between individual correlations between the accuracy and complexity measures (Table 46) show that accuracy and grammatical complexity were found to be negatively correlated at matching units in the within-individual correlations. AS unit level accuracy (A1) with length of AS unit (C1) had a moderate negative correlation (r = -.300); and clause level accuracy (A2) and clause length (C2) had a moderate negative correlation (r = -.357). These level-specific relationships were negligible in the group level correlations, (r = -.073, r = -.123, respectively). And, as described in the complexity results section, the complexity measure was a significant time-varying predictor in the related accuracy model. However, as discussed, this negative correlation was unsurprising given the calculation of the accuracy measures as percentage of error-free units.

The within-individual correlations show that clauses per AS unit (C3) was negatively correlated with AS unit accuracy (A1) (r = -.216) but positively correlated with clause accuracy (A2) (r = .299). The between-individual correlation echo the pattern (r = -.088 and r = .302, respectively). This means that increased subordination (C3) was correlated with decreased AS unit level accuracy (A1) but increased clause level accuracy (A2). These findings are also unsurprising since an increase in the number of clauses in an AS unit mathematically increases the chance that there will be a error in at least one, but an increase in the number of clauses can

increase the proportion of accurate clauses. As such, rather than a tension between control and risk-taking, this negative correlation is more a function of the calculation of the accuracy measurements.

Table 46 Correlations between the Complexity and the Accuracy Measures

	Error-free AS units (A1)	Error-free Clauses (A2)
Within-individual Correlations		
Length of AS unit (C1)	-.3003**	.1392**
Clause Length (C2)	-.1263*	-.3574**
Clauses/AS unit (C3)	-.2160**	.2994**
Lexical Variety (C4)	-.0100	.1253*
Between-individual Correlations		
Length of AS unit (C1)	-.0729	.2500*
Clause Length (C2)	-.0133	-.1229
Clauses/AS unit (C3)	-.0877	.3016**
Lexical Variety (C4)	.2410*	.3144**
*p < .05 **p < .01		

Further, the results also showed that the level 2 predictors of initial proficiency and language background were more influential than the tension between length of unit and accuracy. Considering the mathematics of calculating accuracy as error-free clauses, the positive within-individual correlations between clause-level accuracy (A2) and the complexity measures, and greater influence of language background and initial proficiency, the results did not support a trade-off between accuracy and complexity.

5.4.1.2 Accuracy and Lexical Variety

The hypothesis of a positive correlation between lexical variety (C4) and accuracy was confirmed in the between-individual correlations with AS unit level accuracy (A1) (r =.241) and with clause level accuracy (A2) (r = .314). At the within-individual level, there was a non-significant relationship (r = -.010) between lexical variety (C4) and AS unit accuracy (A1) while the lexical variety (C4) had a significant positive correlation (r = .125) with clause-level

accuracy. As such, an increase in lexical variety was not correlated with a decrease in accuracy, as would be the case if students make errors when stretching to use a variety of words. Rather, the results indicated that overall there was a positive correlation between lexical variety and accuracy, specifically at the clause level. This finding is interesting because it would be expected any possible negative impact from using a less-than fully-learned word would be found at the clause level.

5.4.2 Accuracy and Fluency

I hypothesized a strong negative correlation between accuracy and fluency because of theoretical and methodological reasons. Skehan predicts a tension between meaning (fluency) and form (accuracy) which leads to a trade-off effect. Also, in this task, the students were given little planning time, forcing the students to do online planning, which should induce trade-off effects.

The within-individual and between-individual correlations between accuracy, as measured by error-free AS units (A1), and the fluency measures (Table 47) list negligible magnitudes. None of the correlations between A1 and the three fluency measures were significant. More interestingly, the correlations between error-free clauses (A2) and the fluency measures were contrary to a trade-off effect. Clause level accuracy (A2) and phonation time ratio (F1) have a weak but positive correlation at the within-individual (r = .148) and between-individual (r = .250) level. Clause level accuracy (A2) and mean length of pause (F2), showed a negative relationship within-individual (r = -.108) and between-individual (r = -.123). Remembering that a decrease in pausing is actually improvement, an increase in clause level accuracy (A2) correlated with an increase in fluency. These findings echo the correlations found in Mizera (2006). Likewise, clause level accuracy (A2) was positively correlated with mean

length of fluent run (F3), within individuals (r = .143) and between-individual (r = .302). An increase in clause-level accuracy correlated with an increase in fluency, in each of the three fluency measures.

Table 47 Correlations between the Accuracy and the Fluency Measures

	Error-free AS units (A1)	Error-free Clauses (A2)
Within-individual Correlations		
Phonation Time Ratio (F1)	-.0438	.1478**
Mean Length of Pause (F2)	.0715	-.1075*
Mean Length of Fluent Run (F3)	.0123	.1426**
Between-individual Correlations		
Phonation Time Ratio (F1)	-.0729	.2500*
Mean Length of Pause (F2)	-.0133	-.1229
Mean Length of Fluent Run (F3)	-.0877	.3016**
*p<.05 **p<.01		

5.4.3 Complexity and Fluency

I hypothesized positive correlations between the grammatical complexity measures and the fluency measures. A negative correlation was predicted between lexical variety (C4) and fluency because of retrieval costs of a varied vocabulary. Table 48 lists the within-individual and between-individual correlations among the complexity and fluency measures.

5.4.3.1 Grammatical Complexity and Fluency

The most general complexity measure, length of AS unit (C1) was correlated with increased fluency. Length of AS unit (C1) and phonation time ratio (F1) had a moderate positive within-individual correlation (r = .373) and between-individual correlation (r = .444). Length of AS unit (C1) and mean length of pause (F2) were negatively correlated (improvements in both measures) in both the within-individual correlation (r = -.363) and between-individual correlation (r = -

.302). Length of AS unit (C1) was positively correlated with mean length of fluent run (F3), only weakly in the within-individual correlation (*r* =.147), but moderately strongly in the between-individual correlation (*r* = .570). Clauses per AS unit (C3) had a similar pattern of correlations with the fluency measures, positive with phonation time ratio (F1) and mean length of fluent run (F3) and negative with mean length of pause (F2), which is equivalent to positive correlations with improved fluency. The correlations between clause length (C2) and the fluency measures were negligible, save for clause length (C2) and mean length of pause (F2) at the within-individual level (*r* = -.102) which indicates both improvement in both measures. Overall, higher grammatical complexity scores were correlated with higher fluency.

Table 48 Correlations between the Complexity and the Fluency Measures

	Phonation Time Ratio(F1)	Mean Length of Pause (F2)	Mean Length of Fluent Run (F3)
Within-individual Correlations			
Length of AS unit (C1)	.3730**	-.3625**	.1465**
Clause Length (C2)	.0842	-.1023*	-.0558
Clauses/AS unit (C3)	.3002**	-.2740**	.1671**
Lexical Variety (C4)	.1613**	-.2306**	.0585
Between-individual Correlations			
Length of AS unit (C1)	.4438**	-.3016**	.5697**
Clause Length (C2)	.0749	-.0417	.0031
Clauses/AS unit (C3)	.4187**	-.2977*	.5596**
Lexical Variety (C4)	.2724*	-.2627*	.3812**
*p<.05 **p<.01			

5.4.3.2 Lexical Variety and Fluency

The correlations with lexical variety (C4) and the fluency measures are similarly positive as well, contrary to my hypothesis. Lexical variety (C4) and phonation time ratio (F1) had a weak, but positive within-individual (*r* = .161) and between-individual correlation (*r* = .272). Lexical variety (C4) and mean length of pause (F2) had a negative correlation (*r* = -.231 and *r* = -.263,

respectively), which means improvement in both constructs. Although there were significant correlations between lexical variety and fluency breakdown (pausing), lexical variety and fluency proceduralization were less connected. Lexical variety (C4) and mean length of fluent run (F3) showed a moderate, positive correlation only at the between-individual level (r = .381), and importantly, not at the within-individual level (r = .059).

5.4.3.3 Summary of Complexity and Fluency

Overall, there were no trade-off effects between complexity and fluency. An increase in grammatical complexity (length or subordination) or lexical variety was most often correlated with an improvement in fluency. Even clause length (C2), which did not correlated with phonation time ratio (F1) or mean length of fluent run (F3), showed a weak correlation with mean length of pause (F2) in that an increase in clause length correlated with a decrease in pausing.

Most unexpected is that lexical variety (C4) is correlated with higher fluency, especially with (the lack of) pausing fluency. This analysis seemed to be connected growers in these data.

5.4.4 Correlation Summary

My third research question was concerned, in part, with the relationships between constructs. There were negative correlations between accuracy and structural complexity, but this finding is interpreted as a function of the impact of the length of the unit in an error-free measure. Clause-level accuracy (A2) was correlated with improvements in fluency, i.e., decreasing pausing and increasing length of fluent run. Improvement in fluency was also correlated with increased

grammatical complexity. Mean length of pause (F2) showed robust findings in that pausing decreased with an increase in length of AS unit, length of clause, and clauses per AS unit.

Lexical variety and accuracy at the clause level (A2) was positively correlated. In addition, lexical variety and grammatical complexity, measured by length of AS unit and clauses per AS unit, were positively correlated. Moreover, lexical variety correlates with more fluent speech, as measured by phonation time ratio (F1) and mean length of pause (F2). As such, it seems that an increase in lexical variety (C4) does not hinder accuracy (at the clause level), grammatical complexity, or fluency. In addition, the HLM results showed that an increase in lexical variety corresponded to an increase in fluency (as measured by phonation time ratio and by mean length of pause). This finding may indicate that lexical variety serves as another measure of general proficiency. Since a decrease in pausing is correlated with an increase in lexical variety, these findings may support a model (e.g., MacWhinney, 2001) where lexical retrieval occurs before the construction of the syntactic frame (rather than during formulation).

6.0 SUMMARY AND GENERAL DISCUSSION

This research was the first comprehensive longitudinal study to investigate the development of complexity, accuracy, and fluency (CAF) in language performance. Specifically, this research showed the development of CAF constructs by using nine measures over several time points in spoken language performance of sixty-six L2 English learners. All measures showed growth, which answered my first research question. The conditioned hierarchical linear models determined which individual differences (based on demographic information) predicted individual growth patterns, which addressed my second research question. The models with time-varying predictors and the correlation analysis addressed my third research question about the relationships between the measures. In addition, the data analysis methodology in this dissertation clarifies the source of variance in the scores (individual variation in language performance and between-individual variation) which allows for better interpretation of the results.

6.1 SUMMARY OF THE GROWTH MODELS

Table 49 summarizes the HLM best-fitting models, listing the measures by construct, the mean initial score for the average-aged (25.3 years) and average initial proficiency (19.2 on the listening placement test) student, the growth trajectory type, any predictors (in initial scores and

slope). These results answer RQ2, “do individual differences explain individual growth?” Table 49 also lists the relevant time-varying predictors (i.e., the other dependent measures) that improve the model. This information partially answers RQ3, regarding the relationships between constructs.

6.1.1 Growth Trajectories

Both of the accuracy measures (percentage of error-free AS units and percentage of error-free clauses), all three grammatical complexity measures (length of AS unit, clause length, and clauses/AS unit) and two of the fluency measures (phonation time ratio and mean length of pause) showed linear change trajectories. Lexical variety had a non-linear trajectory, showing a slight initial decline and followed by steeper increase, while the non-linear trajectory of mean length of fluent run (F3) showed steep growth which slowed over time.

Interestingly, there was no other finding of different change trajectories in the measures. For instance, even though lower proficiency students have lower initial scores (for most measures), the lower proficiency students improve in language performance at the same rate as the higher proficiency students.

6.1.2 Predictors

Initial proficiency was, expectedly, the most common predictor of initial scores. Initial proficiency explained variance in initial scores for seven measures, with only clause length (C2) and clauses/AS unit (C3) not differing by initial proficiency. Higher proficiency predicted better performance for each of these measures. Language background was only predictive for three of

the nine measures; both accuracy measures and one fluency measure (mean length of fluent run). In all three measures, students with an Arabic language background outperformed students with a non-Arabic background. This finding is unexplained without a review of the specific error-types. Age was also predictor for mean length of fluent run (F3) in which longer stretches of speech between pauses was found as age decreased. Instruction cohort also explained some of the variance in lexical variety (C4) scores. Finally, gender (with initial proficiency) explained the variance in length of AS unit (C1) scores, and gender also explained a difference in change rate in length of AS unit (C1) scores. The female students had lower initial scores, but that deficiency in length of AS unit was eliminated by a steeper growth rate after time in the intensive English program.

6.1.3 Explanatory Power

Proportion of variance explained gives information about how useful the predictors are in explaining the variance found in the data. Unlike the reporting of effect sizes with other statistical analysis, there is no standardized scale for evaluating the usefulness of HLM predictors (Roberts & Monaco, 2009). In these data, the independent variables explained much of the existing variance. Notably, gender and initial proficiency explained 80% of the variance in the initial scores and gender explained 64.4% of the variance in the slope of the mean length of AS unit (C1). Instruction cohort and initial proficiency explained 46% of the variance in initial score of lexical variety. For the fluency measures, initial proficiency explained 22% of the variance in the initial scores of phonation time ratio and 16% of the initial scores of mean length of pause while initial proficiency, language background, and age explained 44.9% of the variance in initial score of mean length of fluent run.

Less impressive proportions in variance were explained for the accuracy measures. Initial proficiency and language background only explained 5% of the variance in initial scores of percentage of error-free AS units (A1) and only 7.5% of the initial scores of error-free clauses (A2). It is important to note that these measures did not have much variance in initial scores to be explained. The unconditioned model for both accuracy measures (A1 and A2) only had a variance component of .010. It might be more relevant to consider what the predictors in the model mean in terms of typical scores. The average (25.3 years old with a placement score of 19.2) Arabic student was expected to have initial clause level accuracy (A2) of 61%. Considering the range of the scores on the placement test (9-27), the lowest scoring (Arabic) student is expected to have an initial clause accuracy score of 48% and the highest scoring (Arabic) student is expected to have an initial score of 71.4%. The typical Arabic student (placement score of 16) is estimated to have an initial score of 57.1% in clause accuracy (A2), the typical Chinese student (placement test of 22), 58.1%, while the typical Korean student (placement score of 20), 55.5%. Teachers and researchers can evaluate if the differences in the predicted initial scores are relevant.

As stated, the variance component remained significant for four measures, even after all predictors were added or tested in the model. There was significant remaining variation in the model of lexical variety for the initial scores and in the models for all three fluency measures in the initial scores and in the change rate. This remaining variance indicates that there seems to be missing explanatory variable(s). A possible explanation for lexical variety variance is (inconsistent) topic effects. A plausible explanation for the remaining variance in the fleuncy measures is extraversion (Ockey, 2011).

Table 49 Summary of HLM Best-fitting Model for Each Measure

Construct measure	initial score	time-invariant predictor - intercept		variance explaind	trajectory		time-invariant predictor slope			variance explained	time varying predictor		remaining variance
Accuracy													
percentage of error-free AS units (A1)	.644	proficiency	+.012	5%	linear	-.100	--			--	C1	-.018	--
		L1	-.074										
percentage of error-free clauses (A2)	.872	proficiency	+.013	7.5%	linear	+.031	--			--	C2	-045	--
		L1	-.068										
Grammatical Complexity													
length (words) of AS unit (C1)	10.36	proficiency	+.252	80%	linear		gen-der	male	5.276	64.2%	--		--
		gender	-.954					female	+6.037				
clause length (C2) (words)	5.892	--		--	linear	+.786	--			--	--		--
clauses/AS unit (C3)	1.737	--		--	linear	+.779	--			--	--		--
Lexical Variety													
lexical variety (C4) (measured by D)	43.01	proficiency	+.984	46%	quad-ratic	-24.73 +63.34	--			--	--		initial score --
		cohort	+6.19										
Fluency													
phonation time ratio (F1)	.600	proficiency	+.008	22%	linear	.223	--			--	C4	+.001	initial score change rate
mean length of pause (F2)	1.163	proficiency	-.017	16%	linear	-.609	--			--	C4	+.003	initial score change rate
mean length (syllables) of fluent run (F3)	4.521	proficiency	+.100	44.9%	quad-ratic	+2.17 -2.44	--			--	--		initial score change rate
		L1	-6.26										
		age	-.053										

Note: Proficiency means (centered) initial proficiency upon enrollment per the in-house listening placement test; age is centered; L1 (Arabic and Non-Arabic), gender and instruction cohort are categorical. The baseline student average-aged (25.3 years), average proficiency Arabic males from cohort 1.

6.1.4 Summary of the Relationships among Measures

This study explored the relationships between the CAF measurements to answer the third research question. As expected, within each construct (Table 50) some measures were highly correlated, which indicates that perhaps both are not required to capture language performance, generally speaking. The two accuracy measures (A1 and A2) were strongly correlated. Thus, for many studies, clause level accuracy might suffice. Phonation time ratio (F1) and mean length of pause (F2) were highly correlated; both capturing fluency breakdown. The general recommendation is to calculate mean length of pause (to capture fluency breakdown) and mean length of fluent run (to capture fluency proceduralization).

Length of AS unit (C1) and clause length (C3) were strongly correlated with each other, but a recommendation was made to calculate complexity by subordination differently. Therefore, research on grammatical complexity might benefit from a minimum of three measures: length of AS unit, length of clause, and ratio of finite clauses to AS units. A measure of general lexical variety (C4) is also recommended for research on language performance.

Negative relationships between constructs was expected but not generally found. Length of unit scores affected the accuracy scores (of error-free units) and controlling for the length of the unit better predicted the trajectory of the accuracy measures. Controlling for length of unit showed that the percentage of error-free AS units does indeed increase with time in the program. Lexical variety (C4) was a relevant predictor in the trajectories of both phonation time ratio (F1) and mean length of pause (F2), in that students with *higher* lexical variety (C4) scores had *higher* fluency measures. Other correlations across constructs were weak to moderate. In particular,

none of three fluency measures were correlated (within-individual or between-individual) with AS unit level accuracy (A1). Likewise, the fluency measures had little relationship with clause length (C2). A general discussion of the broad implications of the research follows.

Table 50 Within-Individual and Between-Individual Correlations for all CAF Measures

Within-Individual Correlations									
	A1	A2	C1	C2	C3	C4	F1	F2	F3
Percentage Error-free AS units (A1)	--								
Percentage Error-free clause (A2)	0.6995**	--							
Length of AS unit (C1)	-0.3003**	0.1392**	--						
Clause Length (C2)	-0.1263*	-0.3574**	0.1221*	--					
Clauses/AS unit (C3)	-0.2160**	0.2994**	0.8365**	-0.4169**	--				
Lexical Variety (C4)	-0.0100	0.1253*	0.2014**	-0.0533	0.1897**	--			
Phonation Time Ratio (F1)	-0.0438	0.1478**	0.3730**	0.0842	0.3002**	0.1613**	--		
Mean Length of Pause (F2)	0.0715	-0.1075*	-0.3625**	-0.1023*	-0.2740**	-0.2306**	-0.8307**	--	
Mean Length of Fluent Run (F3)	0.0123	0.1426**	0.1465**	-0.0558	0.1671**	0.0585	0.4919**	-0.2409**	--
Between-Individual Correlations									
	A1	A2	C1	C2	C3	C4	F1	F2	F3
Percentage Error-free AS units (A1)	--								
Percentage Error-free clause (A2)	0.8763**	--							
Length of AS unit (C1)	-0.0729	0.2500*	--						
Clause Length (C2)	-0.0133	-0.1229	0.2960**	--					
Clauses/AS unit (C3)	-0.0877	0.3016**	0.8727**	-0.1917	--				
Lexical Variety (C4)	0.2410*	0.3144**	0.3096**	0.1967	0.2093*	--			
Phonation Time Ratio (F1)	-0.1164	0.0551	0.4438**	0.0749	0.4187**	0.2724*	--		
Mean Length of Pause (F2)	0.0485	-0.0826	-0.3016**	-0.0417	-0.2977*	-0.2627*	-0.8401**	--	
Mean Length of Fluent Run (F3)	0.0979	0.2849*	0.5697**	0.0031	0.5596**	0.3812**	0.5014**	-0.3227**	--

6.2 GENERAL DISCUSSION

In the following sections, the implications of the major findings are explored, specifically the lack of different paths of development (Section 2.1) and the lack of trade-off effects (Section 2.2), and the implications of choice of measurements (Section 2.3).

6.2.1 Paths of Development

This study showed that language performance develops in each of the measures of each construct, answering the first research question. The results from this study show that there is indeed intra-individual variation which supports a dynamic systems theory of development. In this study of 66 participants, however, for many measures (accuracy, clause length, clauses/AS unit, and lexical variety), there was no significant variation in trajectories to even investigate if students take multiple paths of development. The change trajectory of eight of the nine measurements did not differ among the students. (The only exception, length of AS unit, the slope for female students was still similar, only steeper which offset a lower initial score.).

6.2.1.1 Shared Developmental Path

This finding of single paths to development was contrary to Larsen-Freeman's (2006) longitudinal case study of repeated written texts which suggested that there may be "preferred paths", specifically a focus on either grammatical complexity (subordination) or on lexical variety. Larsen-Freeman reports results on written texts, not spoken performance, and used

different measures of grammatical complexity (measured by clauses per t-unit) and lexical variety (measured by the number of word types divided by the square root of two times the number of words) than the current study. As mentioned in the literature review, Larsen-Freeman's study only included five students, and although the group mean increased over time for each measure, the difference in mean scores was not statistically significant. A closer look at the estimated scores indicates that three of the five participants showed some growth in both grammatical complexity and lexical variety, while one participant had very similar scores (i.e., no change) across all four observations for both measures, and one participant had no change for one measures and very little change in the other. Although it is possible to surmise the results of these five students indicate multiple paths to development, there are other explanations (e.g., insensitive measures, the time frame was too short to allow development, the within-individual scores simply vary).

In order to look for individual paths, Larsen-Freemen converted the observed scores to z-scores by replacing the raw score with the scores distance from that individual's mean score on that measure. As such, it was often the case that two of the four observations for each measure were above the mean while two observations were below the mean. A review of the graphs seems to show that lexical variety and grammatical complexity plotted on opposite sides of the mean only seven times of the possible twenty (four observations from five students) observations. Therefore, it is not clear that grammatical complexity and lexical variety are separate "attractors".

Larsen-Freeman also plotted the raw scores for the grammatical complexity and lexical variety on a single graph, and it seemed to show that one student made more growth in lexical variety while the other seem to develop more along the grammatical complexity axis. Larsen-

Freeman did not report correlations (specifically within-individual correlations) on the measures of interest, but with estimates (of the raw scores plotted on a given graph), it seems like grammatical complexity and lexical variety are positively correlated for four of the five participants, including the one student singled-out by Larsen-Freeman as focused on lexical variety, and negative for one participant (who in fact showed little growth in either measure). As such, another interpretation of Larsen-Freeman's data might support that grammatical complexity and lexical variety are actually connected growers, for at least most of her learners, as was found in this study.

As stated, the current study found that clauses per AS unit (C3) and lexical variety (C4) were significantly positively correlated in both the within-individual and between-individual correlations. Since the within-individual correlations were positive and significant at the .001 level, the results did not indicate that students must choose to focus their development on one at the expense of the other, but rather these two constructs grew together.

There was remaining variation in the fluency measures, but predictors in the study, even time-varying predictors (i.e., the other scores on the measures), were not relevant. Consequently, even though the variation remains unexplained, the findings did not support a separate path explanation.

In particular, the lack of variation in change trajectories is contrary to the "rich get richer effect" (Stanovich, 1986) in which more proficient students improve more quickly. For one measure, *less* proficient students had improved more quickly, as was found in Wendel's (2007) research. (This dissertation found the less proficient participants (females) had a steeper gain in length of AS unit, C1; Wendel reported no difference in words per sentence in the written texts

but a significant gain in clauses per sentence.) In general, these results showed that all students improve similarly, not that "the rich get richer".

6.2.1.2 Communication Success and Control

The lack of variation in the change trajectories was also contrary to Higgs and Clifford's (1982) suggestion that students with relatively high fluency and vocabulary do not progress in accuracy. Higgs and Clifford's (1982, p. 73) concern from "experiential but consistent data" that communication success, (i.e. getting your idea across) inhibits the need to produce grammatically accurate language was not supported by these data. It must be pointed out that Higgs and Clifford suggested a longitudinal study of participants already in the seeming terminal 2/2+ stage (only scoring 2 or 2+ out of 5 on language proficiency), but that suggested population might be biased. As Hammond (1988, p. 408) emphasizes, the Higgs and Clifford's article lacks "direct empirical evidence" and needs to be tested. This dissertation tests the hypothesis, specifically by determining how vocabulary and fluency are related to accuracy during development.

Higgs and Clifford's hypothesis is directional, i.e., sufficient competency in vocabulary and fluency inhibits continued growth in accuracy. This HLM analysis of both of the accuracy measures (A1 and A2) showed extremely homogenous change trajectories ($p > .500$ and $p = .476$). All students showed similar growth trajectories in accuracy. For percentage of error-free AS units (A1), accuracy showed improvement when controlling for length of AS unit. Percentage of error-free clauses (A2) showed growth over time, with a steeper slope when controlling for length of clause. Importantly, all students showed the same pattern, rather than some students reaching a plateau in development. Moreover, In order to specifically check for a negative effect of fluency on accuracy, the time-varying phonation time ratio (F1) was tested in

each accuracy model, and it was found to not be a predictor. Evidence against Higgs and Clifford's hypothesis was also found when considering lexical variety scores. Lexical variety (C4) had a small positive relationship with clause level accuracy (A2) with the within-individual correlation (r =.125), which was statistically significant, and no correlation was found at the AS unit level (r = -.01). (The between-individual correlations showed significant positive correlations with both accuracy measures, but as stated, within-individual correlations are more valid when considering trade-off effects within language acquisition.) Therefore, using this quantitative analysis, high fluency and/or high lexical variety did not negatively impact accuracy.

In an attempt to explain the difference in conclusions, I ran simple bivariate Pearson correlations between measures on individuals' scores. These correlations were run on the accuracy scores, fluency scores, and lexical variety scores of all individual using the multiple observations of the individual. Since these Pearson correlations do not adjust for the lack of independence between observations, they violate an important assumption and are used only in attempt to explain the difference in conclusions.

Since the correlations between accuracy and lexical variety were generally positively correlated, even this microanalysis seems to indicate that high vocabulary skills do not hinder accuracy growth. On the other hand, the individual correlations between accuracy and the fluency measures were mixed positive and negative. A look at these unreliable correlations might explain the impression that fluency hinders accuracy, as suggested by Higgs and Clifford's (1982) grammatical accuracy hypothesis. More students had a negative (but insignificant) correlation between accuracy at the AS level and phonation time ratio (F1), and more students had a positive (albeit insignificant) correlation with mean length of pause (F2). In other words, some students had better fluency (measured by phonation time ratio and pause length) but lower

accuracy (measured in error-free AS units), and some students had lower fluency and higher accuracy. These results could lead to the impression that fluency inhibits growth in accuracy. It is plausible that poor sentence level accuracy and fluency stand out to teachers or evaluators amid so much variation. Fluent students who are grammatically inaccurate might be conspicuous whereas fluent and accurate students are less memorable. Since Higgs and Clifford (1982, p. 70) asserted their hypothesis based on "experience", these unbalanced student performances might have skewed their impressions of development.

To be clear, this sub-pattern was not found with clause level accuracy (A2), which seemed to be a better measure of accuracy in this study. Twice as many students had a positive (albeit insignificant) correlation between A2 and phonation time ratio (F1) than students with a negative correlation. More students had a negative correlation with mean length of pause (F2) and A2 (which is improvement in both) and more students had a positive correlation with mean length of fluent run (F3) and A2. As such, clause-level accuracy cannot even anecdotally explain why "experience shows" (Higgs & Clifford, 1982, p. 74) that fluency hinders accuracy. All in all, these empirical data do not support the hypothesis that communicative competence (high fluency and/or vocabulary) hinders further growth in accuracy.

6.2.2 Trade-off effects

The current data did not show trade-off effects between constructs as was found in cross-sectional research, summarized in Table 1 and repeated here as Table 51. This was a surprising finding because trade-off effects are commonly found in the literature, especially a fluency-accuracy trade-off. In this section, I discuss possible reasons for the difference in findings: differences in coding and measurements, differences in analysis, and difference in designs. These

longitudinal data suggest that individual development does not show any trade-off effects that have been found in cross-sectional research.

Table 51 Empirical Findings Showing Trade-off or "Competitive" Effects

researcher(s)	design	task	participants	trade-off effects
Ahmadian & Tavakoli (2011)	between groups; 1-way ANOVA	oral narrative about video	intermediate English L2, Persian L1, adult females (n = 60)	accuracy (error-free clauses; verb forms) vs. fluency(# of syllables/min. of speech =PTR; pruned PTR) with COLP
				complexity (subordination; syntactic variety) vs. fluency with COLP
Yuan & Ellis (2003)	between groups; 1-way ANOVA	oral narrative about cartoon	English L2, Chinese L1 undergraduates (n =42)	accuracy vs. fluency with COLP
				accuracy vs. lexical complexity with OLP
Michel, Kuiken & Vedder (2007)	2 X 2 (+/- few elements, +/-mono)	oral info. sharing task	intermediate Dutch L2 from Turkey and Morocco (L1s not given) (n = 44)	accuracy vs. fluency (only in combined monologic and dialogic conditions)
Skehan & Foster (1997)	2 X 2 (planned/ unplanned; post-task/ no post-task	oral task (personal information, narrative, decision-making task)	pre-intermediate English L2, mixed L1 adult (n = 40)	accuracy (proportion of error-free clauses vs. complexity (clauses/c-units)
Skehan (2009a)	between group	various oral tasks	low-intermediate English L2	lexical complexity (D) vs. grammatical complexity -subordination
				lexical complexity (P_Lex) vs. accuracy (error-free clauses)
				lexical complexity (P_Lex) vs. grammatical complexity -subordination

6.2.2.1 Different Coding and Measurements

One possible explanation for this noteworthy difference in results is a difference in coding and measurement. For instance, Skehan (2009a) reported a positive correlation between lexical variety as measured by D and accuracy, which matches the current findings in the correlations at the between-individual level. The current research also found a positive within-individual correlation, but only at the clause level accuracy measure. When Skehan found a negative

correlation between accuracy and lexical variety, it was using P_Lex to measure lexical variety. As mentioned in the literature review, the text-external lexical measures, such as P_Lex, have complications which have not been resolved and are especially difficult to compare across topics, and across studies. Likewise, when Yuan and Ellis (2003) found a negative correlation between accuracy and lexical variety, they measured lexical variety by mean segmental type-token ratio. Thus, results from the different lexical variety measures simply may not be comparable.

In the same way, a difference in clause coding might be a factor in the different findings about the relationship between grammatical complexity and lexical variety. Skehan reported a negative correlation with lexical variety and complexity by subordination, whereas the current data found significant positive correlations at both the within-individual and between-individual correlations. However, as discussed earlier, the complexity by subordination measure (C3) includes finite and non-finite clauses, and it is unclear how other studies defined "clause" but generally complexity by subordination means subordinate finite clauses. It is likely that measurement differences explain some of the differences in results, but I propose that coding and measures differences are not the only explanation for the lack of expected trade-off effects.

6.2.2.2 Different Data Analysis

Previous research, for the most part, reports results of aggregated data. With aggregated data, any "trade-off" effects found could be a result of some students focusing on one aspect of the language performance and other students focusing on another. Skehan (2009b) specifically offered this possibility when his proposed trade-off effects are not found. In fact, separating the correlations to within-individual and between-individual correlations is much stronger than aggregating data. Actual trade-off effects should be detected at the within-individual level, since trade-off effects are hypothesized to be exerted *within* the individual. Additionally, aggregating

data tends to inflate correlations (Ostroff, 1993), which could lead to apparently significant correlations that would not be found in non-aggregated data. The artificial inflation of results from the statistical method, however, does not elucidate the change in polarity of some of the correlations.

Before addressing a likely source of the lack of trade-off effects in these data, I will first dispel the possibility that the current results are driven by the development (improvement) in the scores. A main difference between this research and previous research is that this research is longitudinal. So, as to make closer comparisons between the previous research and the current research, separate between-individual correlations were run on the observations at the first four time points. This analysis would be similar to the correlation analysis employed in studies looking at performance status at a single time. Like Mizera (2006), the results show connected growers rather than competitive trade-off effects. Although the correlations do not always reach significance at each time point, the results follow the same pattern as the longitudinal analysis. With this additional analysis of the data, it is clear that the change in scores (development) is not driving the positive correlations. Therefore, even limiting the analysis to between-individual correlations at single time-points, the current research is contrary to previous trade-off effect findings.

6.2.2.3 Different Research Design

The research design and conclusions drawn from cross-sectional research designs are likely the principal explanation for difference in support for trade-off effects. 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. In addition, group performances (aggregated data) are often inferred to represent, do not necessarily reflect,

individual performances. Therefore, caution is required when making conclusions about possible trade-off effects within-individual performance or development.

In some studies which report trade-off effects, the cross-sectional design might have induced a difference in focus during language performance. For instance, Yuan and Ellis (2003) studied the effect of planning on oral language performance and concluded that there was a trade-off effect between accuracy and fluency based on group score comparisons. The trade-off effect was not found *within* each planning group. Similarly, Skehan and Foster (1997) used between task group means as support for trade-off effects, even though trade-off effects were not found *within* each task, which makes the trade-off effect only at the study level, not even at the group level. Importantly, within-individual correlations, which could illuminate if individual students prioritized one construct over another, were not reported in either study.

Ahmadian and Tavakoli (2011) found that accuracy, complexity and fluency can all improve after repetition of an online planning task. As in Skehan and Foster's (2009b) personal information task and Ahmadian and Tavakoli's repetition with online planning task, the current research showed that accuracy, complexity and fluency can all develop. And the current findings show that CAF can develop without a specific focus on careful online planning and without repetition of the same task (with the same topic).

Despite intuitive appeal for the inevitability of trade-off effects, the current study has refuted the major rationalizations that are given to dismiss results that show positive growth in multiple constructs. First, the task was held constant, without being explicitly designed to induce orientation toward accuracy or fluency. It could be argued that although the task was within a larger curriculum and perhaps the current findings are result of a "balanced goal development" (Skehan, 1998a) in the overall curriculum. Although this is possible, the lack of trade-off effects

was confirmed by the correlation of measures at individual time points. Even more persuasively, the more sophisticated multi-level modeling and the within-individual correlations found no support for a trade-off hypothesis in language performance development.

Although it is enlightening research for pedagogy that students in the careful planning group are more accurate but less fluent, these results only indicate what the students are likely to do in a single performance in a strongly induced careful planning situation. Such findings can assist teachers in assigning tasks to encourage practice in one construct over another. These types of studies, however, have limited value in making hypothesis about language development. The current results show that the development of the subsystems of language performance (complexity, accuracy, and fluency) did not come at the expense of another construct, despite variations found in individual performances. And at least some constructs may even be considered connected growers, such as phonation time ratio (F1) and lexical variety (C4) and mean length of pause (F2) and lexical variety (C4).

Eysenck (1981) cited in Dewaele and Furnham (1999) with work on extraversion suggests 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 his speech. This possibility could explain the lack of trade-off effects. If this theory is validated, extraverts have an advantage not only in fluency (assumingly) but in all areas of language development. In any case, parallel-processing offers a theoretical explanation of why trade-off effects are not inevitable. Similarly, that risk-taking in language performance including making errors (Ortega, 2009, p. 40) is a plausible underlying factor of the finding the students with higher fluency may also have higher accuracy.

6.2.3 Implications of the Findings

The advantage of using multi-level analysis methodology with these longitudinal data allowed a more detailed interpretation of the development of language performance. In fact, Larsen-Freeman (2006) suggests that multivariate analysis might be useful to clarify the messiness of language development data. Based on observations from these 66 participants, this HLM analysis did not find evidence for multiple paths of development as suggested by Larsen-Freeman's research study of five learners. Further, the longitudinal design enabled models of *change* rather than models of *status*. As such, the major difference in findings (e.g., the lack of trade-off effects) does not seem as paradoxical as might have first seemed. Instead, the learners in this population had gains in all constructs, and there was support for connected growers, rather than competitive effects. Given the variability in performances, the data analysis used in this dissertation, although new to the field of applied linguistics, can better explore observations of participants within groups and better explore the complexity of language development.

6.2.3.1 Attentional Resources Discussion

As mentioned in Chapter 3, Section 1, trade-off effects were expected because of generally accepted limited attentional resources. These data question the inevitability of trade-off effects, and therefore, some underlying assumptions. The theoretical underpinning of trade-off effects is the limited capacity during online processing. The three theoretical frameworks reviewed (limited attentional capacity model, cognition hypothesis, and dynamic systems theory) accept attentional resources are limited, but differ on the how the resources are used. For Skehan (2008) this limited capacity is from a single-source view of attention. These results cast doubt on a single-source limited attentional capacity model. Moreover, accepting the inevitability of trade-

off effects, Skehan and Foster (2001, p. 193) state that "performance on a particular task can, at most, help *some* of the areas of language development, not all…", and then propose instructional sequences in order to "foster balance in IL (i.e., interlanguage) and performance". In fact, it may be that performance on any particular task cannot help all areas of language development, as supported by the findings based on the recorded speaking activity discussed in this study.

For Robinson, there are still limited attentional resources despite different pools of attentional resources, Robinson's (2007) cognition hypothesis includes resource dispersing factors (which does deplete attentional resources) and resource-directing variables (which do not degrade performance). In Robinson's view, performance is not necessarily hindered because there are separate pools of attentional resources that the learner can draw from. Since these data did not find trade-off effects, it is possible that these data support a multiple pools of attentional resources. However, it is unclear why attentional resources are retrieved from separate pools of attentional resources in the current study since the task was held constant.

For researchers following a dynamic systems view (de Bot, 2008; Van Geert, 2008) although attentional resources are limited, performance may not be hindered because interconnected "connected growers" required fewer attentional resources which means trade-off effects are not necessarily found. The results may support the concepts of "connected growers" from dynamic systems theory. These data found that most measures were connected growers. Thus, every CAF connection is "meaningful" and connected. If all language performance constructs are meaningfully connected, the theory is underspecified in that it only states that all of the constructs of language performance (complexity, lexical variety, accuracy, and fluency) are connected.

6.2.4 Implications for the Measurements of CAF

This research, with multiple measures of the constructs of language performance, complexity, accuracy, and fluency (CAF), also illuminated the relative usefulness of measures used for this research and suggests how each may be better defined. For each construct, general rather than specific measures were used. Each construct is discussed in turn.

6.2.4.1 Accuracy Measures

General measures of accuracy, specifically percentage of error-free units, were used. It was assumed these measures would be less susceptible to L1 influence, because as Schachter (1974) showed in her seminal paper, raw number of errors ignores what was produced correctly and can be deceiving. L1, however, was found to be a significant predictor (with a bigger impact than initial proficiency, and length of unit, and proficiency and length of unit combined) in both accuracy models. It may be a cultural artifact or a L1 influence, but without further study of the data (e.g., a classification of specific error-type), there is no explanation for this finding.

As discussed in Chapter 5 Section 1.4, the larger scale accuracy measure, the percentage of error-free AS units (A1), was especially confounded accuracy with complexity. The participants in this study had difficulty producing completely error-free AS units, especially as the length of the AS unit increased. As such, this measure may be too demanding for learners at this proficiency level. It may be, however, be a valid measure for more proficient participants, written texts, or less impromptu speeches. Minimally, it is important for researchers to consider the impact of using error-free units.

Additionally, these error-free accuracy measures do not capture any changes in the raw number of errors. Bygate (2001) has suggested that the number of errors per AS unit would be

useful measure because it includes all the errors in the utterance, but that measure would also be highly influenced by length of unit. The use of accuracy based on errors/AS units has been used in the field (e.g., Michel, Kuiken, & Vedder, 2007). They found that the complex task condition resulted in higher accuracy. The interpretation of that finding, however, should consider the effect of the length of the AS unit. In fact, the positive effect of complexity on accuracy was driven by the dialogue condition, which had fewer clauses/AS unit and assumingly fewer words/AS unit, though that information was not reported. Therefore, interpretations based on accuracy based on AS unit should consider the likely effect of the length of unit. A measure such as errors per 100 words would control for the length of unit and may be a viable alternative despite the fact that the unit (100 words) has no psychological or linguistic validity.

6.2.4.2 Complexity Measures

In Chapter 5, Section 2.6 I discussed the structural complexity measures and the lexical variety measures of this study. The results, showing that lexical variety patterns differently than the other complexity measures, support Skehan's (2009b) call that language performance should be measured by a lexical measure in addition to the grammatical complexity measures.

None of the grammatical complexity measures, as calculated here, is recommended to serve as a placement test. Specifically, clause length (C2) and clauses/AS unit (C3) did not distinguish students placed at the low-intermediate level from students placed at the high-intermediate level which makes them questionable. The global measure of grammatical complexity, length of AS unit (C1), showed an unpredicted gender effect in initial scores, but did not reflect general ability per se since the difference was neutralized with time in the program.

In addition, two of the grammatical complexity measures, C2 and C3, seemed to interact in unplanned ways because of the equality of non-finite clauses with finite clauses in the coding

system . In order to capture subclausal complexity, clause length (C2) might be better measured (or additionally measured) as "the average length of all *finite* clauses" (emphasis added) (Norris & Ortega, 2009, p. 561). The measure of subordination (C3), using Foster et al.'s definitions for AS units and clauses, was problematic. Their definition of clause included non-finite verbs with a complement or an adjunct. Although this clause definition was sufficiently easy to consistently code (an important consideration), it seemed to misrepresent some forms of complex language. An AS unit of ten words without a non-finite clause would have ten words in that clause while a ten word AS unit with a non-finite clause would be calculated as an average of five words per clause, lowing the mean clause length (C2) score for including a non-finite clause. I expect that this unplanned result of the Foster et al. coding scheme, greatly reduced the sensitivity of the clause length measure (C2). This finding could help researchers interpret findings based on this coding scheme. For instance, using the same coding, De Jong and Vercellotti (2011) reported no difference in phrasal complexity (measured as words/clause) in the language produced in response to five different picture prompts. Any difference might have been obscured if certain prompts encouraged structures with non-finite clauses (e.g., verb complement structures *decide to go home* or *start to run the race*), which in turn reduced the mean words per clause of speeches based on that prompt.

Moreover, this measure of subordination does not represent the generally expected concept of subordination, where a complex sentence consists of an independent finite clause and one or more subordinate finite clauses. Thus, a better measure of subordination would be a ratio of finite clauses per AS unit. As such, a subordinate finite clause would be defined by the presence of a subordinate conjunction, but then the coding would have to adjust for cases where participants introduce both the subordinate clause and the independent clause with a subordinate

conjunction, so that the ratio would not be artificially raised. This coding issue should be resolved, especially since Huang (2010, p. 163) states that English learners with a Chinese language background are prone to this pattern because conjunction pairs (e.g., if…then, because…so) are collocated in Chinese. For Chinese speakers, this pattern does not seem redundant, but it shows balance between the clauses, in that both clauses are dependent on the other, which is favored.

This study also has implications for the question of how complexity of language performance should be defined. Complexity may be defined as extended and elaborated language. Considering the overall findings of the study with little trade-off effect between accuracy, however, complexity in language performance might better be divorced from a definition including the "willingness to experiment" or "upper limit" of language proficiency.

6.2.4.3 Fluency Measures

I found that the fluency measures worked well. As a general measure of speaking time, phonation time ratio is sufficient. While phonation time ratio (F1) gives a nice picture of general fluency, it was the least informative if the other two measures are used. It is highly correlated with mean length of pause (F2), which is unsurprising given that pausing is included in the calculation of phonation time ratio. Phonation time ratio was included to capture fluency differences from a decrease in the number of pauses while the length of pauses stays the same. Considering the findings, an increase in the number of pauses was not common.

The findings, however, suggest that the measures of mean length of pause and mean length of fluent run are more informative. Mean length of pause (F2) and mean length of fluent run (F3) give complementary information about the fluency of the language performance. Mean length of pause captures any change (e.g., decline) in the length of pauses between utterances,

which indicates how long the speaker pauses before speaking while mean length of fluent run captures any change (e.g., growth) in the length of the utterances between pauses of 200ms or more, which indicates how much speech the speaker can produce before a pause. Both of these measures are recommended for research with oral data. Given these two measures, the single measure of phonation time ratio, which was highly correlated with mean length of pause, seemed unnecessary.

7.0 CONCLUSIONS

7.1 SUMMARY AND DISCUSSION

This research found a lack of individual paths in development and a lack of trade-off effects between the constructs of language performance, complexity, accuracy, and fluency. These two major findings have several theoretical implications. First, these longitudinal data did not offer support for alternative paths in language development; the findings did not find that students take markedly different paths in development. Trajectories were the same (save for length of AS unit) even when initial scores differed significantly. This intensive English program appeared to benefit all students equally. Although incoming proficiency affected initial scores, initial proficiency did not impact the change trajectory. Therefore, the results did not find that more proficient students have steeper gains (Stanovich, 1986; Wendel, 2007) nor a fluency handicap as described by Higgs and Clifford (1982). Second, the accepted limited attentional resources did not result in trade-off effects in language performance during these topic-centered monologues. Trade-off effects were not even found within observations with correlations run at individual time points. Although there is individual variation in scores, the variation was explained for the most part by individual differences (e.g., initial proficiency, language background, gender, cohort, age). For both accuracy measures and the four complexity measures, there was no significant remaining variation in the data to be explained (e.g., by individual paths). Even

though predictors explained a substantial percentage of the variance of fluency measure scores, variance remained in the both the initial scores and in the change rate, but the remaining variance was not explained by the performance other measures.

This research offers data analysis, HLM and within- and between correlations, new to the field of applied linguistics that can be used to examine longitudinal data (observations) of participants within groups. Considering the variability in performances, these statistical methods can better explore the complexity of language development. In addition, this research offers a recommendation about which measurements seem to be less successful for similar research: percentage of error free AS units, complexity by subordination which includes non-finite clauses, and phonation time ratio. As such, the findings have pedagogical implications for proficiency testing and program evaluation. One pedagogical implication of these finding is that vocabulary instruction may foster fluency development.

7.2 LIMITATIONS

This study was limited to the proficiency range in this intensive English program. It would be especially useful to have additional observations as development continues after attendance in the intensive English program. Although this study's observations spanned three academic semesters (over nearly a year) which is notably longer than most studies in the field, an even longer study may possibly find different results, such as multiple trajectories, non-linear trajectories, and different slope trajectories. This study was also limited to students learning English at a single intensive English program.

Although hierarchical linear modeling can test for non-linear models, forcing the data into any particular form may be objectionable within dynamic systems theories. Moreover, the results are limited to the specific measure of accuracy, complexity, and fluency used in the study.

7.3 FUTURE RESEARCH

7.3.1 Populations

This research should be replicated with different populations, specifically populations with other language backgrounds, populations from other intensive English programs, and with students studying English as a foreign language in a non-immersion setting. Most notably, this population heavily consisted of Arabic L1 speakers. Although the analysis was capable of separating language background as a time-invariant predictor, the language backgrounds were not evenly distributed. A study with students with shared language background but from several countries might be able to separate L1 effects from cultural effects, which could not be done in this study. In addition, the research should be replicated in other intensive English programs since the overall finding of growth in all constructs could be a result of this program's curriculum. Also, the findings may be different with students who are not in an immersion program in an English language environment. For instance Ahmadian (2011) reported substantially lower scores in each of the three shared measures (percentage of error-free clauses, length of AS unit, and clauses/AS unit) than reported here in his English as a foreign language in a non-immersion setting.

7.3.2 Measures

Although the nine different dependent variables were carefully chosen to cover much of the constructs, other or additional measures could have given a more complete or even different picture. Another measure of complexity by subordination might well show another pattern, perhaps of a leveling off of growth with higher proficiency. Likewise, a measure of accuracy based on a count of errors may reflect a different pattern, which is needed to more closely explore the L1 differences found in the accuracy measures. A more detailed analysis of the errors in the data, including comparing accuracy based on number of errors to accuracy based on error-free units, may be better able to revisit Schachter's (1974) work on the impact of avoidance on accuracy in language production. An in-depth look at the errors might show that the type of errors (e.g., lexical, syntactic, morphological) can give a richer explanation of the findings, as was found by Kuiken and Vedder (2007). It is possible that the language differences may be found in a particular error-type or that the type of error students make changes over time.

7.3.3 Related Studies

A follow-up study is warranted to test the effects of topic, since the current research indicated that topic effects may affect lexical variety scores in unpredictable ways, especially if comparisons across cohorts are important. As Chalhoub-Deville (2001) pointed out, researchers must not assume that all elicitation prompts are equal. A more detailed comparison of topic effects would also have pedagogical applications. In addition, it would be useful to compare the language performance of the ungraded introductory RSA sessions to see if the performances differ significantly during ungraded tasks.

This study found a relationship between lexical variety and fluency. This relationship should be explored further. In particular, it would be interesting to more directly investigate how fluency (or fluency breakdown) during performance can give insight to the theoretical question of when lexical retrieval occurs during language processing.

Finally, more research is needed to explain the remaining variance in the fluency measures. Perhaps independent variables, either time-invariant (e.g., assertiveness) or time-varying (e.g., interest in the topic) might explain additional variance in fluency scores.

Overall, these finding make a considerable addition to the field, particularly the finding of uniform, linear growth of most measures, with no support of trade-off effects in development.

APPENDIX A

TOPIC PROMPTS

1. Childhood Meal (cohort 1; low-intermediate)
   Describe your favorite meal from childhood. What are the ingredients for this dish? Who made it for you?
2. Transportation (cohort 1; low-intermediate)
   Compare the transportation in your country to the transportation in Pittsburgh.
3. Admired Person (cohort 1; low-intermediate)
   Talk about a famous person whom you admire.
4. Best Friend (cohort 2; low-intermediate)
   Describe your best friend from childhood. How did you meet? What qualities help describe your friend? What did you used to do together?
5. A Surprise (cohort 2; low-intermediate)
   Talk about a day when someone or something surprised you. When did this happen?
6. Vacation Spot (Cohort 1; high-intermediate)
   Talk about your ideal vacation spot. What will you do there?
7. Renting (cohort 1; high-intermediate)
   Talk about renting an apartment, either in your country or in Pittsburgh.
8. Home City (cohort 2; high-intermediate)
   Describe the city you come from. Where is it? How big is it? What kinds of things can you do there? Are there lots of parks?
9. Job (cohort 2; high-intermediate)
   Describe a job that you would love to have. What are the expectations of this job? What are the things that you would love about the job?
10. Disliked Custom (cohort 2; high-intermediate)
    Describe a custom, in you culture, or in another culture, which you do not like. Give details about the expectations of the custom, and describe the things that you don't like about it.

11. Famous Person (cohort 2; high-intermediate)

    Talk about a famous person from the past. This person could be from your country or from another country. Who was this person, and why was he famous?

12. World Problem (cohort 1; high-intermediate)

    Describe a problem in the world that concerns you.

13. A Regret (cohort 1; high-intermediate)

    Talk about something you regret that you have done. What should you have done?

14. Media Violence (cohort 1; advanced)

    Talk about media violence in your country. In your opinion, should violence on tv be banned?

15. Computerized Society (cohort 1; advanced)

    Describe the advantages and disadvantage of living in a computerized society.

16. Extravagant Lifestyle (cohort 1; advanced)

    Describe an extravagant lifestyle. Compare and contrast an extravagant lifestyle with an ordinary lifestyle.

17. Rich and Poor (cohort 1; advanced)

    When the gap between the rich and the poor is so large, you need to balance a desire for luxury with compassion for the needs of others. Do you agree or disagree? Why?

18. Internet Risks (cohort 1; advanced)

    What are some of the risks of using the internet? How can you protect yourself?

APPENDIX B

PARTICIPANTS' DEMOGRAPHIC INFORMATION

Table 52 Participants' Demographic Information

participant	cohort	sex	age	proficiency	entered at	language background	semesters	speeches
848	2	m	32	16	low-inter	Arabic	2	5
948	2	f	22	16	high-inter	Arabic	1	3
1060	1	m	19	14	low-inter	Arabic	3	7
1061	1	m	20	24	low-inter	Arabic	3	7
1062	1	m	19	16	low-inter	Arabic	3	7
1067	2	m	22	11	low-inter	Arabic	2	6
1073	1	m	21	9	low-inter	Arabic	3	7
1075	1	m	27	15	low-inter	Arabic	3	7
1077	1	f	33	16	low-inter	Arabic	2	4
1081	1	m	32	12	low-inter	Arabic	3	7
1085	1	f	25	13	low-inter	Chinese	3	7
1110	1	f	28	16	low-inter	Arabic	3	7
1117	2	f	28	20	high-inter	Chinese	1	3
1118	1	f	30	16	low-inter	Chinese	2	5
1147	2	m	26	16	low-inter	Arabic	2	5
1150	1	m	21	23	high-inter	Arabic	2	4
1152	2	f	23	19	low-inter	Arabic	2	5
1153	2	m	21	17	low-inter	Arabic	2	5
1155	2	m	26	13	low-inter	Arabic	2	5
1156	1	m	23	24	high-inter	Arabic	2	4
1158	2	f	26	11	low-inter	Arabic	2	5
1159	2	m	29	14	low-inter	Arabic	2	5
1160	1	m	21	20	high-inter	Arabic	2	4
1162	1	m	20	21	high-inter	Arabic	2	4
1163	2	m	34	11	low-inter	Arabic	2	5
1164	2	m	30	17	low-inter	Arabic	2	5
1165	1	m	23	26	high-inter	Arabic	2	4

participant	cohort	sex	age	proficiency	entered at	language background	semesters	speeches
1166	1	m	27	21	high-inter	Arabic	2	4
1167	2	m	25	17	low-inter	Arabic	2	5
1168	1	f	22	23	high-inter	Arabic	2	4
1169	2	m	24	18	low-inter	Arabic	2	3
1172	1	f	25	22	high-inter	Chinese	2	4
1182	1	m	23	23	high-inter	Arabic	2	4
1183	2	m	24	18	low-inter	Arabic	2	5
1187	1	f	25	24	high-inter	Chinese	2	4
1188	2	m	29	16	low-inter	Chinese	2	5
1189	1	f	33	24	high-inter	Chinese	2	3
1193	2	m	24	12	low-inter	Arabic	2	5
1199	1	f	28	22	high-inter	Chinese	2	4
1208	1	f	22	21	high-inter	Korean	2	4
1211	1	m	30	15	high-inter	Arabic	2	4
1212	2	f	25	12	low-inter	Arabic	2	5
1214	2	m	25	13	low-inter	Arabic	2	5
1245	2	f	25	24	high-inter	Arabic	1	3
1247	2	m	18	27	high-inter	Arabic	1	4
1254	2	f	23	19	high-inter	Arabic	1	4
1258	2	f	23	23	high-inter	Arabic	1	3
1260	2	f	19	23	high-inter	Arabic	1	4
1263	2	m	19	24	high-inter	Arabic	1	4
1275	2	f	33	23	high-inter	Korean	1	3
1285	2	f	35	22	high-inter	Korean	1	4
1286	2	f	31	25	high-inter	Chinese	1	3
1287	2	f	25	25	high-inter	Chinese	1	3
1289	2	f	34	22	high-inter	Chinese	1	3
1293	2	f	19	23	high-inter	Korean	1	4
1295	2	m	30	21	high-inter	Korean	1	4
1299	2	m	25	20	high-inter	Korean	1	4
1300	2	m	26	24	high-inter	Chinese	1	3
1305	2	f	21	24	high-inter	Chinese	1	3
1311	2	f	29	20	high-inter	Chinese	1	3
1325	2	f	29	21	high-inter	Arabic	1	4
1327	2	f	29	21	high-inter	Arabic	1	4
1334	2	f	26	22	high-inter	Chinese	1	4
1337	2	f	26	23	high-inter	Chinese	1	3
1343	2	f	28	20	high-inter	Korean	1	3
1351	2	m	18	25	high-inter	Arabic	1	4

APPENDIX C

SCORES PER OBSERVATION

Table 53 Scores for Each Measure per Observation

participant	ELI_dur_fract_year	% error-free AS units (A1)	% error-free clauses (A2)	length (words) AS unit (C1)	clause length (words) (C2)	clauses/ AS unit (C3)	lexical variety(C4)	phonation time ratio (F1)	mean length of pause (F2)	mean length of fluent run (F3)
848	0.0778	0.2778	0.3636	7.667	6.273	1.222	50.16	0.5645	1.082	3.980
848	0.1611	0.3333	0.2778	9.444	4.722	2.000	28.07	0.5689	1.171	4.107
848	0.3806	0.4286	0.4737	8.357	6.158	1.357	29.53	0.5506	1.234	3.933
848	0.4667	0.5455	0.6500	9.909	5.450	1.818	45.66	0.6273	0.942	3.549
848	0.5139	0.3636	0.4118	9.818	6.353	1.545	43.26	0.6038	0.891	3.377
948	0.0556	0.7500	0.7222	10.500	7.000	1.500	62.58	0.5856	1.002	4.082
948	0.1444	0.7273	0.7826	12.636	6.043	2.091	73.33	0.6573	0.713	4.389
948	0.1972	0.4545	0.7143	11.273	5.905	1.909	61.04	0.6858	0.720	3.843
1060	0.1056	0.3077	0.6071	13.615	6.321	2.154	79.14	0.6584	0.745	5.352
1060	0.1778	0.3333	0.6667	12.222	6.111	2.000	42.27	0.4839	1.387	4.816
1060	0.2250	0.3000	0.4286	8.950	6.393	1.400	33.38	0.7756	0.829	6.524
1060	0.4056	0.2000	0.5938	13.200	6.188	2.133	60.20	0.7552	0.577	5.807
1060	0.4722	0.5000	0.6923	13.000	5.000	2.600	37.07	0.6079	1.130	4.905
1060	0.7222	0.1818	0.5000	14.000	7.700	1.818	67.44	0.7126	0.577	3.908
1060	0.8333	0.3750	0.6500	15.750	6.300	2.500	77.42	0.8131	0.567	4.480
1061	0.0639	0.3077	0.6667	12.769	5.030	2.538	42.80	0.5838	1.104	4.896
1061	0.1222	0.4762	0.6571	9.190	5.514	1.667	46.43	0.6228	0.938	5.019
1061	0.1889	0.5385	0.7000	12.077	5.233	2.308	34.01	0.5443	1.035	4.481
1061	0.4139	0.4286	0.5556	11.714	9.111	1.286	56.15	0.4179	1.415	2.900
1061	0.4944	0.3333	0.5385	14.000	6.462	2.167	51.57	0.4500	1.448	4.163
1061	0.7583	0.3846	0.5833	12.000	6.500	1.846	62.15	0.6348	0.797	4.596
1061	0.8556	0.2000	0.1538	12.300	9.462	1.300	46.42	0.5880	1.046	4.234
1062	0.1056	0.2143	0.3810	8.571	5.714	1.500	35.17	0.7016	0.716	4.073

participant	ELI_dur_fract_year	% error-free AS units (A1)	% error-free clauses (A2)	length (words) AS unit (C1)	clause length (words) (C2)	clauses/ AS unit (C3)	lexical variety(C4)	phonation time ratio (F1)	mean length of pause (F2)	mean length of fluent run (F3)
1062	0.1778	0.2941	0.3636	8.588	6.636	1.294	37.73	0.6883	0.793	4.582
1062	0.2250	0.4000	0.5517	7.650	5.276	1.450	45.03	0.6650	0.671	4.339
1062	0.4056	0.4286	0.6522	10.214	6.217	1.643	39.31	0.6716	0.585	3.657
1062	0.4722	0.5000	0.7778	12.250	5.444	2.250	49.10	0.6156	0.796	4.063
1062	0.7306	0.1250	0.3600	13.875	4.440	3.125	63.78	0.7698	0.546	4.860
1062	0.8361	0.1250	0.5263	14.750	6.211	2.375	84.83	0.6420	0.661	4.215
1067	0.1056	0.5000	0.4583	9.063	6.042	1.500	51.60	0.6138	0.886	4.034
1067	0.1778	0.2778	0.4091	9.278	7.591	1.222	30.10	0.5579	1.104	5.543
1067	0.2250	0.4000	0.6250	8.600	5.375	1.600	40.00	0.5384	1.094	4.580
1067	0.7111	0.4783	0.5200	6.000	5.520	1.087	36.51	0.5572	1.066	4.113
1067	0.7972	0.2500	0.6333	15.333	6.133	2.500	42.36	0.6629	0.748	4.571
1067	0.8444	0.3000	0.4286	10.100	7.214	1.400	37.52	0.5132	1.071	3.769
1073	0.1056	0.5294	0.5455	8.118	6.273	1.294	45.85	0.6572	0.917	4.319
1073	0.1778	0.3077	0.5217	9.231	5.217	1.769	22.80	0.4934	1.312	4.825
1073	0.2250	0.3125	0.4667	9.563	5.100	1.875	40.91	0.5211	1.174	4.792
1073	0.4083	0.4286	0.5517	11.857	5.724	2.071	46.14	0.6544	0.762	4.518
1073	0.4917	0.5000	0.6000	9.313	4.967	1.875	50.84	0.6056	0.902	4.333
1073	0.7222	0.2667	0.5000	10.533	5.267	2.000	67.34	0.6530	0.765	5.729
1073	0.8333	0.4167	0.4500	9.417	5.650	1.667	44.00	0.5220	1.014	4.057
1075	0.1056	0.5882	0.7143	10.588	6.429	1.647	56.19	0.7505	0.673	5.358
1075	0.1861	0.6667	0.8235	13.889	7.353	1.889	31.58	0.5744	1.022	5.041
1075	0.2250	0.1111	0.4762	14.111	6.048	2.333	66.92	0.6157	0.832	4.098
1075	0.4056	0.1875	0.4839	11.063	5.710	1.938	54.49	0.5653	0.901	5.017
1075	0.4722	0.3333	0.5833	11.000	5.500	2.000	37.33	0.4345	1.228	3.063
1075	0.7222	0.3000	0.5000	10.000	6.250	1.600	34.98	0.6057	0.917	4.581
1075	0.8333	0.5714	0.8125	13.000	5.688	2.286	105.8	0.5929	0.929	3.698
1077	0.1056	0.5385	0.6111	9.077	6.556	1.385	45.38	0.4721	1.139	3.389
1077	0.1778	0.8462	0.8462	7.385	7.385	1.000	29.50	0.5540	1.129	3.510
1077	0.2250	0.7692	0.8235	7.462	5.706	1.308	40.85	0.5346	1.260	4.311
1077	0.4139	0.7273	0.8235	10.364	6.706	1.545	92.18	0.5144	1.089	3.764
1077	0.4722	0.5833	0.7059	10.167	7.176	1.417	32.15	0.5030	1.209	3.706
1077	0.7306	0.4167	0.5500	11.667	7.000	1.667	47.57	0.6120	0.872	4.093
1077	0.8528	0.7273	0.8889	14.818	6.037	2.455	69.60	0.6943	0.593	3.969
1081	0.1056	0.1667	0.5000	9.833	4.917	2.000	29.21	0.7411	0.956	2.656
1081	0.1778	0.3333	0.4545	8.222	6.727	1.222	26.65	0.6146	1.378	3.366
1081	0.2250	0.0000	0.1250	10.667	8.000	1.333	41.65	0.5787	1.337	3.349
1081	0.4139	0.0000	0.2222	12.500	8.333	1.500	32.17	0.6774	1.219	3.571
1081	0.5056	0.0000	0.3846	16.200	6.231	2.600	32.69	0.6783	1.051	3.500
1081	0.7222	0.3333	0.3333	13.667	6.833	2.000	47.33	0.6737	0.796	2.942

particip ant	ELI_dur_fr act_year	% error-free AS units (A1)	% error-free clauses (A2)	length (words) AS unit (C1)	clause length (words) (C2)	clauses/ AS unit (C3)	lexical variety(C4)	phonation time ratio (F1)	mean length of pause (F2)	mean length of fluent run (F3)
1081	0.8333	0.0000	0.2632	11.778	5.579	2.111	70.53	0.7405	0.717	3.259
1085	0.1056	0.4545	0.5000	8.182	7.500	1.091	33.10	0.5250	0.940	2.636
1085	0.1778	0.2500	0.2222	8.125	7.222	1.125	18.94	0.4152	1.382	2.952
1085	0.2250	0.6154	0.5789	6.692	4.579	1.462	66.30	0.4850	1.086	2.868
1085	0.4139	0.4545	0.6667	9.000	4.125	2.182	49.32	0.6005	0.872	3.653
1085	0.5056	0.1429	0.5556	14.286	5.556	2.571	30.45	0.5386	1.095	2.902
1085	0.7306	0.2500	0.3571	12.625	7.214	1.750	33.03	0.6945	0.638	3.393
1085	0.8361	0.3750	0.6818	16.625	6.045	2.750	47.23	0.6676	0.604	3.542
1110	0.1056	0.2727	0.2500	9.545	6.563	1.455	58.56	0.4997	1.325	3.634
1110	0.1778	0.5385	0.6111	9.769	7.056	1.385	39.95	0.6365	1.002	5.125
1110	0.2250	0.4615	0.5556	8.231	5.944	1.385	47.90	0.5750	0.978	4.368
1110	0.4139	0.2857	0.4400	16.286	4.560	3.571	46.90	0.6335	0.838	4.255
1110	0.4722	0.1429	0.2222	10.429	8.111	1.286	22.15	0.5772	0.843	3.514
1110	0.7222	0.0769	0.4000	14.769	6.400	2.308	63.39	0.7944	0.615	7.429
1110	0.8333	0.3333	0.5000	25.500	6.955	3.667	69.98	0.7626	0.591	5.080
1117	0.0556	0.2500	0.3889	9.500	6.333	1.500	51.14	0.6742	0.740	3.580
1117	0.1417	0.1667	0.5000	11.250	5.192	2.167	53.47	0.7200	0.510	3.350
1117	0.1889	0.1818	0.3889	10.545	6.444	1.636	52.08	0.6859	0.593	3.431
1118	0.1056	0.5000	0.6000	6.250	5.000	1.250	34.37	0.4512	1.367	2.848
1118	0.1778	0.3333	0.4000	6.778	6.100	1.111	27.73	0.4033	1.863	2.722
1118	0.2250	0.3636	0.5000	7.818	5.375	1.455	32.23	0.4392	1.605	3.300
1118	0.4139	0.3000	0.6316	9.900	5.211	1.900	38.20	0.4886	1.242	3.292
1118	0.5056	0.8889	0.8333	11.889	5.944	2.000	26.29	0.5187	1.221	3.250
1147	0.0778	0.5000	0.5000	6.714	4.700	1.429	46.22	0.4840	1.662	4.147
1147	0.1611	0.1667	0.3333	7.667	5.111	1.500	50.49	0.5695	1.104	3.956
1147	0.3889	0.5714	0.6538	9.071	4.885	1.857	53.04	0.6058	0.862	4.800
1147	0.4667	0.3333	0.5357	10.400	5.571	1.867	63.63	0.6925	0.767	4.474
1147	0.5222	0.3333	0.6316	11.667	5.526	2.111	32.50	0.6375	0.831	3.722
1150	0.0833	0.4545	0.7143	14.000	5.500	2.545	42.70	0.6100	0.889	5.204
1150	0.1750	0.6000	0.7931	9.067	4.690	1.933	42.77	0.5169	1.730	6.161
1150	0.4000	0.6875	0.8148	10.125	6.000	1.688	42.78	0.6601	0.732	5.385
1150	0.5056	0.5000	0.8000	15.875	6.350	2.500	39.21	0.4876	1.094	3.612
1152	0.0833	0.7778	0.6800	6.667	4.800	1.389	45.49	0.6123	0.919	3.463
1152	0.1639	0.4286	0.5385	8.143	4.385	1.857	49.13	0.6338	0.983	4.400
1152	0.3806	0.0909	0.4400	12.091	5.320	2.273	41.52	0.7213	0.755	3.368
1152	0.4861	0.0909	0.4400	13.455	5.920	2.273	42.13	0.7844	0.644	4.556
1152	0.5222	0.3333	0.5238	11.417	6.524	1.750	55.68	0.8191	0.556	4.407
1153	0.0778	0.5333	0.5714	8.267	5.905	1.400	82.81	0.4737	1.529	5.000
1153	0.1611	0.5556	0.7059	10.167	5.382	1.889	76.22	0.6469	1.031	5.396

particip ant	ELI_dur_fr act_year	% error-free AS units (A1)	% error-free clauses (A2)	length (words) AS unit (C1)	clause length (words) (C2)	clauses/ AS unit (C3)	lexical variety(C4)	phonation time ratio (F1)	mean length of pause (F2)	mean length of fluent run (F3)
1153	0.3806	0.7368	0.7857	9.316	6.321	1.474	72.64	0.6292	0.885	5.113
1153	0.4694	0.4615	0.6452	16.385	6.871	2.385	63.91	0.6549	0.802	4.900
1153	0.5222	0.1818	0.4286	13.545	7.095	1.909	70.38	0.5896	0.699	3.449
1155	0.0833	0.6000	0.7143	7.300	5.214	1.400	44.93	0.4531	1.232	3.029
1155	0.1639	0.7778	0.8333	6.778	5.083	1.333	68.35	0.5891	0.923	3.458
1155	0.3806	0.7333	0.6667	8.200	5.857	1.400	40.81	0.6323	0.684	3.685
1155	0.4694	0.7692	0.8611	14.615	5.278	2.769	55.82	0.7322	0.651	5.216
1155	0.5167	0.7333	0.8000	10.800	8.100	1.333	39.23	0.7126	0.745	5.431
1156	0.0833	0.6250	0.7931	11.563	6.379	1.813	69.84	0.6895	0.776	6.289
1156	0.1639	0.5625	0.7353	11.375	5.353	2.125	58.73	0.6627	0.779	6.244
1156	0.4000	0.5000	0.6923	14.250	6.577	2.167	51.80	0.7164	0.731	6.628
1156	0.5056	0.5000	0.7826	21.750	7.565	2.875	60.89	0.6309	0.845	5.275
1158	0.0833	0.4286	0.5455	8.714	5.545	1.571	61.78	0.3105	2.809	3.286
1158	0.1639	0.1250	0.4000	8.750	4.667	1.875	26.76	0.3640	2.155	3.406
1158	0.3833	0.3636	0.5000	8.000	6.286	1.273	37.19	0.5421	1.592	4.032
1158	0.4694	0.3750	0.5652	15.500	5.391	2.875	35.80	0.6876	0.880	4.659
1158	0.5222	0.4545	0.5625	8.727	6.000	1.455	43.21	0.6342	1.314	3.829
1159	0.0833	0.3846	0.4118	6.846	5.235	1.308	69.21	0.5648	0.893	3.154
1159	0.1639	0.6471	0.6818	6.235	4.818	1.294	63.62	0.4214	1.423	4.000
1159	0.3806	0.4444	0.5238	7.222	6.190	1.167	43.83	0.7226	0.681	4.600
1159	0.4694	0.5714	0.7241	10.857	5.241	2.071	61.79	0.8199	0.574	5.429
1159	0.5222	0.6923	0.7647	10.462	8.000	1.308	59.48	0.7761	0.506	4.388
1160	0.0778	0.3000	0.5909	11.250	5.114	2.200	48.40	0.7439	0.685	7.860
1160	0.1611	0.3571	0.5667	12.429	5.800	2.143	80.33	0.5945	1.222	6.195
1160	0.3917	0.1304	0.4528	11.217	4.868	2.304	64.89	0.7291	0.855	8.171
1160	0.5028	0.2143	0.4828	11.357	5.483	2.071	55.57	0.6390	1.068	7.306
1162	0.0833	0.0000	0.3500	16.200	8.100	2.000	33.30	0.7312	0.724	4.979
1162	0.1639	0.5556	0.7143	9.667	6.214	1.556	42.02	0.5430	1.048	3.500
1162	0.4278	0.2667	0.5357	11.667	6.250	1.867	45.12	0.7351	0.595	5.500
1162	0.5028	0.4615	0.6538	12.385	6.192	2.000	47.98	0.7593	0.532	5.148
1163	0.0778	0.3571	0.4375	6.214	5.438	1.143	38.56	0.3781	1.446	2.612
1163	0.1611	0.2000	0.4583	11.400	4.750	2.400	28.53	0.5507	0.957	2.909
1163	0.3806	0.4286	0.6000	7.714	5.400	1.429	37.82	0.5102	0.953	3.388
1163	0.4694	0.1111	0.3846	9.778	6.769	1.444	34.51	0.4030	1.368	3.089
1163	0.5222	0.3077	0.4444	9.538	6.889	1.385	50.38	0.4760	1.113	3.364
1164	0.0833	0.6250	0.7000	7.125	5.700	1.250	56.97	0.3866	1.990	3.129
1164	0.1639	0.5000	0.6667	10.375	5.533	1.875	45.20	0.5165	1.310	2.596
1164	0.3806	0.7143	0.7500	7.929	6.938	1.143	31.24	0.7476	0.964	3.053
1164	0.4694	0.6250	0.7778	12.125	5.389	2.250	67.29	0.5925	1.103	4.025

particip ant	ELI_dur_fr act_year	% error-free AS units (A1)	% error-free clauses (A2)	length (words) AS unit (C1)	clause length (words) (C2)	clauses/ AS unit (C3)	lexical variety(C4)	phonation time ratio (F1)	mean length of pause (F2)	mean length of fluent run (F3)
1164	0.5222	0.5000	0.6875	10.200	6.375	1.600	47.25	0.6939	0.645	3.000
1165	0.0778	0.4286	0.7273	12.857	5.455	2.357	70.46	0.6875	0.756	5.980
1165	0.1611	0.4000	0.7241	14.800	5.103	2.900	58.42	0.6495	0.860	5.000
1165	0.4000	0.2727	0.6000	14.636	6.440	2.273	58.84	0.7737	0.670	5.961
1165	0.5056	0.5000	0.7241	17.000	5.862	2.900	87.56	0.7032	0.606	4.962
1166	0.0833	0.1250	0.3889	14.375	6.389	2.250	36.77	0.4878	1.258	3.723
1166	0.1639	0.5556	0.7692	14.000	4.846	2.889	44.25	0.6356	0.745	4.172
1166	0.3917	0.2500	0.6471	14.583	5.147	2.833	68.48	0.8082	0.532	5.245
1166	0.5028	0.2857	0.4643	13.857	6.929	2.000	65.90	0.8515	0.496	6.360
1167	0.0833	0.4286	0.6000	9.214	5.160	1.786	63.23	0.6571	0.725	4.457
1167	0.1639	0.4000	0.6538	8.133	4.692	1.733	55.57	0.6018	0.810	3.654
1167	0.3806	0.3333	0.4615	7.778	5.385	1.444	40.35	0.6414	0.720	3.607
1167	0.5167	0.4375	0.5217	9.125	6.348	1.438	40.21	0.7572	0.612	4.480
1167	0.5222	0.2857	0.5000	11.429	6.154	1.857	50.64	0.7505	0.561	4.118
1168	0.0750	0.6000	0.6364	10.333	7.045	1.467	53.95	0.6671	0.674	4.362
1168	0.1417	0.7931	0.8868	9.000	4.925	1.828	50.79	0.7225	0.604	6.377
1168	0.3917	0.5000	0.7826	18.125	6.304	2.875	67.47	0.7268	0.635	8.441
1168	0.5028	0.2000	0.7273	26.800	6.091	4.400	49.23	0.7565	0.502	6.077
1169	0.0778	0.6000	0.6667	10.400	5.778	1.800	60.58	0.7243	0.722	4.580
1169	0.1611	0.3333	0.6667	13.067	5.444	2.400	48.04	0.7168	0.745	6.196
1169	0.3806	0.3333	0.6250	14.583	7.292	2.000	41.40	0.7142	0.665	5.173
1172	0.0750	0.4615	0.5455	8.923	5.273	1.692	40.07	0.4687	1.250	2.882
1172	0.1417	0.1667	0.3000	9.500	5.700	1.667	28.33	0.4995	1.013	3.357
1172	0.3917	0.1333	0.4595	13.400	5.432	2.467	43.36	0.6614	0.781	5.620
1172	0.5056	0.4375	0.6471	10.750	5.059	2.125	41.92	0.6817	0.893	7.333
1182	0.0750	0.3571	0.6667	9.571	4.467	2.143	48.81	0.5378	0.753	3.292
1182	0.1417	0.6667	0.8182	9.083	4.955	1.833	38.56	0.4921	0.974	3.311
1182	0.3917	0.7500	0.8824	12.500	5.882	2.125	45.35	0.5584	0.883	4.774
1182	0.5028	0.2857	0.6364	13.000	5.515	2.357	58.22	0.7417	0.641	5.702
1183	0.0778	0.2500	0.3500	7.750	6.200	1.250	59.66	0.7632	0.570	4.962
1183	0.1611	0.3600	0.5106	9.720	5.170	1.880	46.81	0.8197	0.499	7.795
1183	0.3806	0.2000	0.5185	10.667	5.926	1.800	53.50	0.8194	0.515	5.660
1183	0.4667	0.1111	0.3750	17.111	6.417	2.667	87.62	0.7582	0.516	4.283
1183	0.5139	0.6364	0.6471	10.273	6.647	1.545	42.34	0.7527	0.783	4.178
1187	0.0750	0.1667	0.4545	14.417	5.242	2.750	53.63	0.6734	0.575	4.061
1187	0.1417	0.2727	0.4800	15.636	6.880	2.273	33.72	0.7268	0.692	5.786
1187	0.4583	0.2941	0.5294	13.471	6.735	2.000	61.76	0.8483	0.420	8.047
1187	0.5222	0.2778	0.6316	13.389	6.342	2.111	36.12	0.8328	0.400	7.659
1188	0.0778	0.1667	0.2222	9.083	6.056	1.500	41.48	0.5667	1.146	4.558

particip ant	ELI_dur_fr act_year	% error-free AS units (A1)	% error-free clauses (A2)	length (words) AS unit (C1)	clause length (words) (C2)	clauses/ AS unit (C3)	lexical variety(C4)	phonation time ratio (F1)	mean length of pause (F2)	mean length of fluent run (F3)
1188	0.1611	0.3077	0.4167	11.308	6.125	1.846	80.59	0.7242	0.695	4.179
1188	0.3806	0.4000	0.5385	10.067	5.808	1.733	41.57	0.6625	0.967	5.600
1188	0.4667	0.3333	0.6000	14.444	6.500	2.222	66.44	0.6647	0.629	3.559
1188	0.5139	0.4545	0.5556	11.000	6.722	1.636	46.96	0.6834	0.607	3.313
1189	0.0778	0.8000	0.8947	12.500	6.579	1.900	91.70	0.5556	1.098	4.348
1189	0.1611	0.5455	0.6522	11.455	5.478	2.091	34.18	0.5894	1.014	4.067
1189	0.3917	0.5882	0.7500	11.294	5.333	2.118	55.16	0.7124	0.658	6.120
1193	0.0778	0.2857	0.3750	7.571	6.625	1.143	32.22	0.4283	1.492	3.031
1193	0.1611	0.3333	0.5294	10.000	5.294	1.889	73.95	0.5912	1.054	3.579
1193	0.3806	0.3333	0.4375	9.750	7.313	1.333	44.50	0.5631	1.158	3.720
1193	0.4667	0.2000	0.3125	12.400	7.750	1.600	49.80	0.5407	1.158	4.160
1193	0.5139	0.3333	0.3636	7.111	5.818	1.222	33.54	0.3882	1.900	3.289
1199	0.0833	0.1250	0.3846	12.500	7.692	1.625	90.28	0.5457	0.995	3.292
1199	0.1639	0.4000	0.6667	12.600	5.250	2.400	26.21	0.2715	3.784	4.684
1199	0.3917	0.5556	0.7368	15.778	7.474	2.111	73.94	0.7705	0.716	6.190
1199	0.5028	0.3000	0.6154	16.700	6.423	2.600	59.30	0.6850	0.776	5.213
1208	0.0750	0.0909	0.3333	9.091	5.556	1.636	58.22	0.5558	0.874	3.102
1208	0.1417	0.5333	0.6364	6.333	4.318	1.467	52.28	0.6026	0.917	3.690
1208	0.4667	0.1667	0.2667	9.833	7.867	1.250	56.38	0.5070	1.169	4.848
1208	0.5889	0.2222	0.2143	14.444	9.286	1.556	76.98	0.5942	0.938	4.042
1211	0.0778	0.1429	0.6471	14.857	6.118	2.429	62.61	0.6569	0.885	3.891
1211	0.1611	0.6923	0.7200	10.077	5.240	1.923	46.30	0.6629	0.758	3.906
1211	0.4000	0.3750	0.6667	11.375	4.667	2.438	48.90	0.8375	0.513	5.872
1211	0.5056	0.5000	0.7368	16.250	6.842	2.375	54.05	0.7613	0.643	4.974
1212	0.0778	0.5000	0.4667	8.500	6.800	1.250	62.01	0.4448	1.123	2.966
1212	0.1611	0.4118	0.5217	9.176	6.783	1.353	66.30	0.7013	0.646	4.365
1212	0.3833	0.4762	0.6071	8.905	6.679	1.333	53.68	0.7767	0.633	4.042
1212	0.4694	0.4000	0.7714	19.100	5.457	3.500	69.25	0.7656	0.516	6.106
1212	0.5222	0.6667	0.8500	16.444	7.400	2.222	61.08	0.7632	0.517	5.538
1214	0.0833	0.3333	0.4375	8.167	6.125	1.333	45.65	0.5971	0.959	3.340
1214	0.1639	0.2500	0.5000	11.250	5.625	2.000	33.29	0.5330	1.063	3.070
1214	0.3806	0.7000	0.7143	7.700	5.500	1.400	34.09	0.5575	1.211	4.731
1214	0.4694	0.7143	0.7778	13.714	5.333	2.571	52.92	0.7162	0.785	4.091
1214	0.5222	0.2667	0.4211	9.267	7.316	1.267	39.03	0.7832	0.616	3.333
1245	0.0583	0.4615	0.5556	10.385	7.500	1.385	44.48	0.7050	0.639	4.232
1245	0.1444	0.4000	0.7308	15.500	5.962	2.600	61.66	0.7638	0.544	3.746
1245	0.1972	0.5000	0.5652	9.063	6.304	1.438	62.92	0.7734	0.562	4.245
1247	0.0583	0.5600	0.6207	8.240	7.103	1.160	51.07	0.6128	1.416	4.375
1247	0.1444	0.3529	0.6744	12.176	4.814	2.529	57.44	0.5785	1.063	9.250

participant	ELI_dur_fract_year	% error-free AS units (A1)	% error-free clauses (A2)	length (words) AS unit (C1)	clause length (words) (C2)	clauses/ AS unit (C3)	lexical variety(C4)	phonation time ratio (F1)	mean length of pause (F2)	mean length of fluent run (F3)
1247	0.1611	0.7692	0.8621	14.538	6.517	2.231	55.92	0.6194	0.966	6.792
1247	0.1917	0.6500	0.7073	12.350	6.024	2.050	70.80	0.7134	0.943	7.071
1254	0.0583	0.1538	0.4348	11.308	6.391	1.769	67.81	0.7157	0.820	9.091
1254	0.1444	0.2857	0.5862	11.571	5.586	2.071	60.84	0.6813	0.923	5.128
1254	0.1611	0.2143	0.4828	10.571	5.103	2.071	49.64	0.6791	0.743	5.444
1254	0.1972	0.1667	0.2000	9.333	5.600	1.667	23.96	0.6078	1.154	4.375
1258	0.0556	0.6154	0.7600	9.769	5.080	1.923	86.73	0.6908	0.777	5.029
1258	0.1417	0.1111	0.5714	14.111	6.048	2.333	89.00	0.6629	0.712	3.942
1258	0.1889	0.6000	0.7368	11.700	6.158	1.900	48.44	0.6548	0.654	3.643
1260	0.0583	0.6429	0.7826	10.286	6.261	1.643	48.35	0.5090	1.421	3.544
1260	0.1444	0.8889	0.9556	11.333	4.533	2.500	59.42	0.6066	1.070	5.342
1260	0.1611	0.5263	0.7188	11.158	6.625	1.684	72.86	0.6082	1.056	7.048
1260	0.1972	0.6842	0.7586	9.105	5.966	1.526	69.84	0.5853	1.232	6.681
1263	0.0583	0.8125	0.8095	8.688	6.619	1.313	68.68	0.6316	1.124	6.650
1263	0.1444	0.6000	0.8621	19.300	6.655	2.900	32.19	0.7275	0.915	4.646
1263	0.1611	0.4375	0.6333	10.938	5.833	1.875	37.98	0.7317	0.707	5.578
1263	0.1917	0.6000	0.6667	11.267	6.259	1.800	56.65	0.7460	0.704	4.944
1275	0.0556	0.1875	0.1579	7.313	6.158	1.188	56.07	0.7852	0.558	5.273
1275	0.1417	0.2727	0.4375	10.091	6.938	1.455	69.49	0.6370	0.735	3.640
1275	0.1889	0.3077	0.4118	7.769	5.941	1.308	98.33	0.6474	0.767	3.915
1285	0.0583	0.5333	0.5789	9.533	7.526	1.267	52.44	0.6229	0.934	5.045
1285	0.1611	0.5000	0.7692	11.571	6.231	1.857	42.34	0.6094	0.936	5.682
1285	0.1611	0.3636	0.6429	13.727	5.393	2.545	61.66	0.6377	1.019	5.950
1285	0.1917	0.3571	0.5417	9.214	5.375	1.714	63.30	0.6510	0.863	5.048
1286	0.0583	0.6000	0.6667	8.667	7.222	1.200	29.95	0.6314	0.911	4.950
1286	0.1444	0.6000	0.8261	12.300	5.348	2.300	82.61	0.6420	0.706	3.561
1286	0.1917	0.3333	0.5909	10.833	5.909	1.833	50.74	0.6493	0.615	3.185
1287	0.0556	0.3333	0.6522	12.917	6.739	1.917	70.64	0.7813	0.512	5.273
1287	0.1417	0.6667	0.8205	11.056	5.103	2.167	55.77	0.7412	0.525	4.864
1287	0.1889	0.5333	0.6429	9.600	5.143	1.867	52.44	0.6644	0.725	4.179
1289	0.0556	0.4167	0.6316	10.083	6.368	1.583	62.82	0.6763	0.870	3.977
1289	0.1417	0.4000	0.7308	14.100	5.423	2.600	63.81	0.7629	0.575	4.320
1289	0.1889	0.3333	0.6250	12.250	6.125	2.000	43.74	0.7229	0.601	4.105
1293	0.0583	0.6154	0.6471	8.231	6.294	1.308	54.66	0.4921	1.389	4.265
1293	0.1444	0.5333	0.6538	9.867	5.692	1.733	37.66	0.6524	0.892	5.143
1293	0.1611	0.3636	0.5238	12.273	6.429	1.909	61.62	0.5776	1.066	4.333
1293	0.1917	0.6875	0.8000	7.875	5.040	1.563	49.02	0.5584	0.975	3.804
1295	0.0583	0.4444	0.5238	7.222	6.190	1.167	36.38	0.5514	1.011	3.824
1295	0.1444	0.2143	0.5667	11.571	5.400	2.143	45.44	0.6223	0.649	3.892

participant	ELI_dur_fract_year	% error-free AS units (A1)	% error-free clauses (A2)	length (words) AS unit (C1)	clause length (words) (C2)	clauses/ AS unit (C3)	lexical variety(C4)	phonation time ratio (F1)	mean length of pause (F2)	mean length of fluent run (F3)
1295	0.1611	0.5882	0.6190	9.000	7.286	1.235	52.16	0.6842	0.546	3.937
1295	0.1917	0.3846	0.5000	8.385	6.056	1.385	45.57	0.6383	0.659	3.125
1299	0.0583	0.5000	0.5385	9.125	5.615	1.625	52.72	0.4152	1.582	3.818
1299	0.1444	0.2857	0.5882	15.571	6.412	2.429	58.99	0.4915	1.121	4.000
1299	0.1611	0.5714	0.7692	12.286	6.615	1.857	65.89	0.4423	1.252	4.047
1299	0.1917	0.2222	0.5000	10.889	6.125	1.778	56.65	0.4651	1.120	3.788
1300	0.0861	0.6154	0.7727	11.385	6.727	1.692	69.23	0.5677	0.957	4.813
1300	0.1444	0.6154	0.6429	12.769	5.929	2.154	72.15	0.5860	0.953	5.083
1300	0.1917	0.4286	0.5909	8.286	5.273	1.571	104.2	0.3848	1.647	3.477
1305	0.0639	0.4286	0.6923	9.571	5.154	1.857	50.23	0.6918	0.624	3.808
1305	0.1417	0.2222	0.5882	17.111	4.529	3.778	53.47	0.7832	0.487	4.173
1305	0.1889	0.3077	0.4762	9.615	5.952	1.615	40.74	0.8124	0.453	4.490
1311	0.0972	0.4286	0.5294	8.929	7.353	1.214	50.01	0.7495	0.532	4.755
1311	0.1444	0.6667	0.7576	11.067	5.030	2.200	42.99	0.8167	0.519	6.511
1311	0.1917	0.4167	0.6364	12.250	6.682	1.833	47.59	0.7794	0.581	6.042
1325	0.0583	0.4615	0.5000	10.692	6.950	1.538	37.28	0.6550	0.929	4.176
1325	0.1444	0.4000	0.5000	11.667	5.833	2.000	49.35	0.7791	0.624	5.571
1325	0.1611	0.3750	0.6333	10.750	5.733	1.875	59.53	0.8208	0.479	4.717
1325	0.1972	0.2667	0.3600	11.267	6.760	1.667	53.34	0.7754	0.611	4.527
1327	0.0583	0.4706	0.4737	9.353	8.368	1.118	54.01	0.7206	0.687	4.577
1327	0.1444	0.2500	0.6000	13.750	6.600	2.083	65.62	0.8339	0.413	4.833
1327	0.1611	0.4000	0.5429	12.867	5.514	2.333	65.04	0.8238	0.550	6.391
1327	0.1972	0.3077	0.5600	13.692	7.120	1.923	71.49	0.8644	0.376	6.038
1334	0.0583	0.4545	0.5625	9.727	6.688	1.455	51.11	0.6425	0.920	5.242
1334	0.1444	0.2727	0.5862	10.909	4.138	2.636	34.76	0.6565	0.821	3.773
1334	0.1611	0.4000	0.6818	10.200	4.636	2.200	46.13	0.5805	1.084	3.750
1334	0.1917	0.2222	0.6316	11.778	5.579	2.111	72.20	0.6122	0.911	3.104
1337	0.0583	0.5455	0.6000	8.364	6.133	1.364	55.64	0.4935	1.011	2.647
1337	0.1444	0.2857	0.7000	12.286	4.300	2.857	59.46	0.5418	1.034	3.289
1337	0.1917	0.2727	0.3846	7.091	6.000	1.182	48.22	0.4830	1.073	2.549
1343	0.0556	0.4000	0.3571	11.400	8.143	1.400	62.52	0.5581	0.921	3.558
1343	0.1417	0.5385	0.6818	10.231	6.045	1.692	50.43	0.5525	0.864	3.267
1343	0.1889	0.4615	0.5625	8.923	7.250	1.231	68.58	0.5345	1.036	3.396
1351	0.0583	0.7222	0.8571	9.556	4.914	1.944	66.86	0.7299	0.678	4.630
1351	0.1444	0.4545	0.7333	14.909	5.467	2.727	54.12	0.7654	0.701	4.865
1351	0.1611	0.3846	0.6087	13.000	7.348	1.769	57.57	0.7761	0.735	5.500
1351	0.1917	0.7647	0.8333	9.529	6.750	1.412	62.85	0.7526	0.685	4.623

APPENDIX D

GROUP MEANS (AND CHANGE) BY LEVEL

Table 54 Group Means of First and Last RSA (and Change) by Proficiency Level

	low-intermediate			high-intermediate			advanced (cohort 1 only)		
	begin. mean (SD)	ending mean (SD)	group change	begin. mean	ending mean	group change	begin. mean	ending mean	group change
% Error-free AS units (A1)	0.4277 (.156)	0.3756 (.186)	-0.0521	0.4195 (.199)	0.4357 (.177)	+0.0162	0.3472 (.185)	0.3612 (.167)	+0.0140
% Error-free clauses (A2)	0.5081 (.137)	0.5272 (.156)	+0.0192	0.5764 (.151)	0.6020 (.152)	+0.0255	0.5743 (.156)	0.6109 (.180)	+0.0366
mean length (words) of AS unit (C1)	8.4247 (1.598)	9.2827 (1.885)	+0.8580	10.411 (2.249)	11.083 (2.171)	+0.6725	12.978 (1.99)	15.383 (4.491)	+2.4035
mean clause length (words) (C2)	5.8795 (.692)	5.5175 (.873)	-0.3620	6.3027 (1.003)	6.0821 (.962)	-0.2206	6.1234 (.903)	6.3133 (.929)	+0.1899
clauses/AS (C3)	1.4403 (.260)	1.7017 (.0349)	+0.2614	1.6898 (.461)	1.8536 (.412)	+0.1638	2.1535 (.406)	2.4488 (.658)	+0.2952
lexical variety (C4)	51.135 (14.00)	49.119 (16.05)	-2.0167	52.247 (14.71)	49.822 (15.36)	-2.4249	55.221 (11.71)	60.564 (17.84)	+5.3429
phonation time ratio (F1)	0.5569 (.1262)	0.5865 (.104)	+0.0296	0.6275 (.0953)	0.6231 (.114)	-0.0045	0.7131 (.077)	0.6930 (17.84)	-0.0201
mean length of pause (F2)	1.1524 (.482)	1.0570 (.335)	-0.0954	0.9157 (.266)	0.9380 (.460)	+0.0223	0.6929 (.133)	0.7247 (.210)	+0.0318
mean length of fluent run (F3)	3.7690 (.855)	4.2644 (1.191)	+0.4954	4.4431 (1.102)	4.3224 (1.067)	-0.1207	5.6130 (1.487)	5.0549 (1.297)	-0.5580

BIBLIOGRAPHY

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*, 1-12.

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

Anderson, J. R. (1983). *The architecture of cognition.* Cambridge, MA: Harvard University Press.

Aoun, J. E., Benmamoun, E., & Choueiri, L. (2010). *The syntax of Arabic.* Cambridge: Cambridge University Press.

Bardovi-Harlig, K., & Bofman, T. (1989). Attainment of syntactic and morphological accuracy by advanced language learners. *Studies in Second Language Acquisition*, 17-34.

Baron-Cohen, S., Knickmeyer, R. C., & Belmonte, M. K. (2005). Sex differences in the brain: Implications for explaining autism. *Science, 310 (4)*, 819-823.

Bates, E., Dale, P. S., & Thal, D. (1995). Individual differences and their implications for theories of language development. In P. Fletcher & B. MacWhinney (Eds.), *The Handbook of Child Language* (pp. 96-151). Cambridge: Basil Blackwell Inc.

Boyle, M. H., & Willms, J. D. (2001). Multilevel modelling of hierarchical data in developmental studies. *Journal of Child Psychology and Psychiatry*, 141-162.

Brumfit, C. (1984). *Communicative methodology in language teaching: The roles of fluency and accuracy.* Cambridge: Cambridge University Press.

Bygate, M. (2001). Effects of task repetition on the structure and control of oral language. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching Pedagogic Tasks: Second Language Learning, Teaching and Testing* (pp. 23-48). Essex: Pearson Education Limited.

Cameron, D. (2009). Sex/gender, language and the new biologism. *Applied Linguistics*, 173-192.

Chalhoub-Deville, M. (2001). Task-based assessments: Characteristics and validity evidence. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching Pedagogic Tasks* (pp. 210-228). Essex: Pearson Education Limited.

Chaudron, C. (1988). *Second language classrooms: Research on teaching and learning.* Cambridge, MA: Cambridge University Press.

Chen, J. F., Warden, C. A., & Chang, H.T. (2005). Motivators that do not motivate: The case of Chinese EFL learners and the influence of culture on motivation. *TESOL Quarterly*, 609-633.

Chen, P. (1999). *Modern Chinese: History and Sociolinguistics.* Cambridge: Cambridge University Press.

Cho, B. E. (2004). Issues concerning Korean learners of English: English education in Korea and some common difficulties of Korean students. *The East Asian Learner*, 31-36.

Corrigan, A., Dobson, B., Kellerman, E., Spaan, M., Strowe, L., & Tyma, S. (1979). *Michigan Test of English Language Proficiency (Form Q).* Ann Arbor: Michigan University Press.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment.* Cambridge: CUP.

Crookes, G. (1989). Planning and interlanguage variation. *Studies in Second Language Acquisition*, 367-222.

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

David, A., Myles, F., Rogers, V., & Rule, S. (2009). Lexical development in instructed L2 learners of French: Is there a relationship with morphosyntactic development. In B. Richards, M. H. Daller, D. Malvern, P. Meara, J. Milton, & J. Treffers-Daller, *Vocabulary Studies in First and Second Language Acquisition: The Interface Between Theory and Application* (pp. 147-163). Hampshire: Palgrave Macmillian.

de Bot, K. (2008). Introduction: Second language development as a dynamic process. *Modern Language Journal, 92*, 166-178.

de Bot, K., Lowie, W., & Verspoor, M. (2007). A dynamic systems theory approach to second language acquisiton. *Bilingualism: Language and Cognition*, 7-21.

De Jong, N., & Perfetti, C. A. (2011). Fluency training in the ESL classroom: An experimental study of fluency development and proceduralization. *Language Learning*, 1-36.

De Jong, N., & Vercellotti, M. L. (2011). Norming picture story prompts for second language production research: Fluency, linguistic items, and speakers' Perceptions. *Paper Presented at AAAL*. Chicago, IL.

DeKeyser, R. M. (2007). Study abroad as foreign language practice. In R. M. DeKeyser (Ed.), *Practice in a Second Language: Perspectives from Applied Linguistics and Cognitive Psychology* (pp. 208-226). Cambridge: Cambridge University Press.

Dewaele, J. M., & Furnham, A. (1999). Extraversion: The unloved variable in applied linguisitic research. *Language Learning*, 509-544.

Dörnyei, Z., & Skehan, P. (2003). Individual differences in second language learning. In C. J. Doughty, & M. H. Long (Eds.), *The Handbook of Second Language Acquisition* (pp. 589-630). Malden, MA: Blackwell Publishing Ltd.

Doughty, C. J., & Williams, J. (1998). Issues and terminology. In C. J. Doughty, & J. Williams, *Focus on Form in Classroom Second Language Acquisition* (pp. 1-11). Cambridge: Cambridge University Press.

Elliot, A. J., & Murayama, K. (2008). On the measurement of achievement goals: Critique, illustration, and application. *Journal of Educational Psychology*, 613-628.

Ellis, N. C. (2006). Selective attention and transfer phenomena in L2 acquisition: Contingency, cue competition, salience, interference, overshadowing, blocking, and perceptual learning. *Applied Linguistics, 27 (2)*, 164-194.

Ellis, R. (1994). *The Study of Second Language Acquistion.* Oxford: Oxford University Press.

Ellis, R., & Barkhuizen, G. (2005). *Analysing Learner Language.* New York: Oxford University Press.

English Language Institute at the University of Pittsburgh. (2007). RSA Evalutation and Rubric. Pittsburgh: English Language Institute.

Flahive, D. E., & Snow, B. G. (1980). Measures of syntactic complexity in evaluating ESL compositions. In J. W. Oller, & K. Perkins (Eds.), *Research in Language Testing* (pp. 171-176). Bowley, MA: Newbury House Publishers, Inc.

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

Francis, D. J., Stuebing, K. K., Shaywitz, S. E., & Shaywitz, B. A. (1996). Developmental lag versus deficit model of reading disability: A longitudinal, individual growth curves analysis. *Journal of Educational Psychology*, 3-17.

Halliday, M. A., & Matthiessen, C. (1999). *Construing Experience through Meaning: A Language-based Approach to Cognition.* Cassell.

Hammond, R. M. (1988). Accuracy versus communicative compentency: The acquisition of grammar in the second language classroom. *Hispania*, 408-417.

Higgs, T., & Clifford, R. (1982). The push toward communication. In T. Higgs, *Curriculum Competence and the Foreign Language Teacher.* Skokie, IL: National Textbook Company.

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

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications.* New York: Routledge.

Huang, L. S. (2010). The potential influence of L1 (Chinese) on L2 (English) communication. *ELT Journal*, 155-164.

Hunt, K. W. (1965). *Grammatical Structures Written at Three Grade Levels.* Champaign: NCTE.

Hunt, K. W. (1970). Syntactic maturity in school children and adults. *Monographs of the Society for Research in Child Development*.

Hyltenstam, K., & Abrahamsson, N. (2003). Maturational constraints in SLA. In C. J. Doughty, & M. H. Long (Eds.), *The Handbook of Second Language Aquistion* (pp. 539-588). Malden, MA: Blackwell Publishing Ltd.

Jared, D., Levy, B. A., Cormier, P., & Wade-Woolley, L. (2010). Early predictors of biliteracy development in children in French immersion: A 4-year longitudinal study. *Journal of Educational Psychology*, 1-20.

Koponen, M., & Riggenbach, H. (2000). Overview: Varying perspectives on fluency. In H. Riggenbach, *Perspectives on Fluency* (pp. 5-24). Ann Arbor, MI: University of Michigan Press.

Kuiken, F., & Vedder, I. (2007). Task complexity and measures of linguistic performance in L2 writing. *IRAL, 45*, 261-284.

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*, 590-619.

Larsen-Freeman, D. (2009). Adjusting expectations: The study of complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics, 30*, 579-589.

Larsen-Freeman, D. (2010). Complex, dynamic systems: A new transdisciplinary theme for applied linguistics? *Language Teaching, 45 (2)*, 202-214.

Larsen-Freeman, D., & Long, M. H. (1991). *An Introduction to Second Language Acquisition Research.* Harlow: Longman Group.

Lee, H. H. (1989). *Korean Grammar*. New York: Oxford University Press.

Lesaux, N. K., Rupp, A. A., & Siegel, L. S. (2007). Growth in reading skills of children from diverse linguistic backgrounds: Findings from a 5-year longitudinal study. *Journal of Educational Psychology*, 821-834.

Li-Grining, C. P., Maldonado-Carreno, C., Votrubal-Drzal, E., & Haas, K. (2010). Children's early approaches to learning and academic trajectories through fifth grade. *Developmental Psychology*, 1062-1077.

Lin, H. (2001). *A Grammar of Mandarin Chinese.* Muenchen: Lincom Europa.

Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly*, 36-62.

Luk, Z. P., & Shirai, Y. (2009). Is the acquisition order of grammatical morphemes impervious to L1 knowledge? Evidence from the acquisition of plural -s, articles, and possessive 's. *Language Learning*, 721-754.

Luoma, S. (2004). *Assessing Speaking.* New York: Cambridge University Press.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzying Talk. Volume II: The database, 3rd Edn.* Lawrence Erlbaum.

MacWhinney, B. (2001). The competition model: The input, the context, and the brain. In P. Robinson (Ed.), *Cognition and Second Language Instruction*. New York: Cambridge University Press.

Mahmound, A. (2000). Modern standard Arabic vs. non-standard Arabic: Where do Arab students of EFL transfer from? *Language, Culture and Curriculum*, 126-136.

Malvern, D., & Richards, B. (1997). A new measure of lexical diversity. In A. Ryan, & A. Wray, *Evolving Models of Lanaguge.* (pp. 58-71). Multilingual Matters.

McCormick, D. E., & Vercellotti, M. L. (2009). To err is human to self-correct divine: Examining classroom recorded speaking activity data to support ESL self-correction as noticing. *Paper Presented at AAAL*. Denver, CO.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 30-46.

McKee, G., Malvern, D., & Richards, B. (2000). Measuring vocabulary diversity using dedicated software. *Literary and Linguistic Computing*, 323-337.

Meara, P., & Bell, H. (2001). P_Lex: A simple and effective way of describing the lexical characteristics of short L2 text. *Prospect*, 5-19.

Meisel, J., Clahsen, H., & Pienemann, M. (1981). On determining developmental stages in natural second language acquistion. *Studies in Second Language Acquistion*, 109-135.

Michel, M. C., Kuiken, F., & Vedder, I. (2007). The influence of complexity in monologic versus dialogic tasks in Dutch L2. *IRAL*, 241-259.

Mills, C. (1990). Syntax and the evaluation of college writing: A blind alley. In L. A. Arena (Ed.), *Language Proficiency: Defining, Teaching, and Testing* (pp. 107-119). New York: Plenum Press.

Mizera, G. J. (2006). Working memory and L2 fluency. *Unpublished Doctoral Disseration*. University of Pittsburgh.

Nation, I. S. (1984). *Vocabulary Lists: Words, Affixes, and Stems.* English Language Institue, Victoria University of Wellington.

Norris, J. M., & Ortega, L. (2003). Defining and measuring SLA. In C. Doughty, & M. Long, *Handbook of Second Language Acquistion* (pp. 717-761). Malden, MA: Blackwell Publishing Ltd.

Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics 30 (4)*, 555-578.

Nydell, M. K. (2006). *Understanding Arabs: A Guide for Modern Times.* Boston: Intercultural Press, Inc.

Ockey, G. (2011). Self-consciousness and assertiveness as explanatory variables of L2 oral ability: A latent variable approach. *Language Learning*, 968-989.

Odlin, T. (2003). Cross-linguistic influence. In C. J. Doughty, & M. H. Long (Eds.), *The Handbook of Second Language Acquisition* (pp. 436-486). Malden: Blackwell Publishing.

Ortega, L. (2009). *Understanding Second Language Acquisition.* London: Hodder Education.

Ortega, L., & Iberri-Shea, G. (2005). Longitudinal research in second language acquisition: Recent trends and future directions. *Annual Review of Applied Linguistics,* 26-45.

Ostroff, C. (1993). Comparing correlations based on individual-level and aggregated data. *Journal of Applied Psychology*, 569-582.

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

Parker, M. D., & Brorson, K. (2005). A comparative study between mean length of utterance in morphemes (MLUm) and mean length of utterance in words (MLUw). *First Language*, 365-376.

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

Politzer, R. L., & McGroarty, M. (1985). An exploratory study of learning behaviors and their relationship to gains in linguistic and communicative competence. *TESOL Quarterly*, 103-123.

Purpura, J. E. (2004). *Assessing Grammar.* New York: Cambridge University Press.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods.* Thousand Oaks: Sage Publications, Inc.

Read, J. (2000). *Assessing Vocabulary.* Cambridge University Press.

Roberts, J. K., & Monaco, J. P. (2009). Effect size measures for the two-level linear multilevel model. *Paper Presented at the American Educatinal Research Association*. Houston, TX.

Robinson, P. (2001a). Task complexity, cognitive resources, and syllabus design: A triadic framework for examining task influences on SLA. In P. Robinson, *Cognition and Second Language Instruction* (pp. 287-318). Cambridge: Cambridge University Press.

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

Robinson, P. (2003). Attention and memory during SLA. In C. J. Doughty, & M. H. Long (Eds.), *The Handbook of Second Language Acquistion* (pp. 631-678). Malden, MA: Blackwell Publishing.

Robinson, P., & Gilabert, R. (2007). Task complexity, the cognition hypothesis and second language learning and performance. *IRAL*, 161-176.

Robinson, P., Cardierno, T., & Shirai, Y. (2009). Time and motion: Measuring the effects of the conceptual demands of tasks on second language speech production. *Applied Linguistics*, 533-554.

Romaine, S. (2003). Variation. In C. J. Doughty, & M. H. Long (Eds.), *The Handbook of Second Language Acquistion* (pp. 409-435). Malden: Blackwell Publishing.

Sawaki, Y., Stricker, L. J., & Oranje, A. H. (2009). Factor structure of the TOEFL internet-based test. *Lanugage Testing*, 5-30.

Scarborough, H. S. (1990). Index of productive syntax. *Applied Pycholinguistics, 11*, 1-22.

Schachter, J. (1974). An error in error analysis. *Language Learning*, 205-214.

Schachter, J., & Celce-Murcia, M. (1977). Some reservations concerning error analysis. *TESOL Quarterly*, 441-451.

Schmidt, R. (1992). Psychological mechanisms underlying second langugage fluency. *SSLA*, 357-385.

Scott, M. S., & Tucker, R. G. (1974). Error analysis and English-learning strategies of Arab students. *Language Learning*, 69-97.

Shafer, D. (2006). *Revolution: Software at the Speed of Thought.* Monterey, CA: Shafer Media.

Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchial models, and individual growth models. *Journal of Educational and Behavioral Statistics*, 323-355.

Singer, J. D., & Willett, J. B. (2003). *Applied Longitudinal Data Analysis.* New York: Oxford University Press.

Skehan, P. (1998a). *A Cognitive Approach to Language Learning.* Oxford: Oxford University Press.

Skehan, P. (1998b). Task-based instruction. *Annual Review of Applied Linguistics*, 268-286.

Skehan, P. (2001). Tasks and language performance assessment. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching Pedagogical Tasks, Second Language Learning, Teaching and Testing* (pp. 167-185). New York: Pearson Education Limited.

Skehan, P. (2009a). Lexical performance by native and non-native speakers on language-learning tasks. In B. Richards, M. H. Daller, D. D. Malvern, P. Meara, J. Milton, & J. Trefferes-Daller, *Vocabulary Studies in First and Second Langauge Acquisition: The Interface Between Theory and Application* (pp. 107-124). Hampshire: Palgrave Macmillian.

Skehan, P. (2009b). Modelling second langauge performance: Integrating complexity, accuracy, fluency, lexis. *Applied Linguistics*, 1-23.

Skehan, P., & Foster, P. (1997). Task type and task processing conditions as influences on foreign language performance. *Language Teaching Research, 1 (3)*, 185-211.

Skehan, P., & Foster, P. (1999). The influence of task structure and processing conditions on narrative retellings. *Language Learning*, 93-120.

Skehan, P., & Foster, P. (2001). Cognition and Tasks. In P. Robinson (Ed.), *Cognition and Second Language Instruction* (pp. 183-205). Cambridge: Cambridge University Press.

Skehan, P., & Foster, P. (2008). Complexity, accuracy, fluency and lexis in task-based performace: A meta-analysis of the Ealing research. In S. Van Dael, A. Housen, F. Kuiken, M. Pierrard, & V. I. (eds), *Complexity, Accuracy, and Fluency in Second Language Use, Learning, and Teaching.* University of Brussels Press.

Sohn, H.-M. (1999). *The Korean Language.* New York: Cambridge University Press.

Spinner, P. (2011). Second language assessment and morphosyntactic development. *Studies in Second Language Acquisition*, 529-561.

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

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly, 21 (4)*, 360-407.

Strauss, L. C., & Volkwein, J. F. (2002). Comparing student performance and growth in 2- and 4-year institutions. *Research in Higher Education*, 133-161.

Syed, Z. (2003). The Sociocultural context of English language teaching in the Gulf. *TESOL Quarterly*, 337-341.

Thompson-Panos, K., & Thomas-Ruzic, M. (1983). The least you should know about Arabic: Implication of the ESL writing instructor. *TESOL Quarterly*, 609-623.

Towell, R., Hawkins, R., & Bazergui, N. (1996). The development of fluency in advanced learners of French. *Applied Linguistics*, 84-119.

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, M. H. Daller, D. D. Malvern, P. Mera, J. Milton, & J. Treffers-Daller, *Vocabulary Studies in First and Second Language Acquisition: The Interface Between Theory and Application* (pp. 74-90). Hampshire: Palgrave Macmillan.

Umbach, P. D. (2007). Gender equity in the academic labor market: An analysis of academic disciplines. *Research in Higher Education*, 169-192.

Van de Pol, M., & Verhulst, S. (2006). Age-dependent traits: A new statistical model to separate within- and between-individual effects. *American Society of Naturalists*, 766-773.

Van Geert, P. (2008). The dynamic systems approach in the study of L1 and L2 acquistion: An introduction. *Modern Language Journal, 92*, 179-199.

VanPatten, B. (2007). Input processing in adult second language acquistion. In B. VanPatten, & J. Williams, *Theories in Second Language Acquisition* (pp. 115-136). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Verspoor, M., Lowie, W., & Van Dijk, M. (2008). Variability in second language development from a dynamic systems perspective. *Modern Language Journal, 92*, 214-231.

Wallentin, M. (2009). Putative sex differences in verbal abilities and language cortex: A critical review. *Brain and Language*, 175-183.

Wang, Z. (2009). Modeling L2 speech production and performance: Evidence from five types of planning and two task structures. *Unpublished Doctoral Dissertation*. The Chinese University of Hong Kong.

Wells, G. (1986). Variation in child language. In P. Fletcher, & M. Garman, *Language Acquisition: Studies in First Language Development* (pp. 109-139). Cambridge: Cambridge University Press.

Wendel, J. N. (2007). An assessment of English language learner writing. *Journal of Bunkyo Gakuin Univerity Department of Foreign Languages and Bunkyo Gakuin College*, 13-41.

Wolfe-Quintero, K., Inagaki, S., & Kim, H. Y. (1998). *Second language development in writing: Measures of fluency, accuracy, and complexity.* University of Hawai'i, Second Language Teaching and Curriculum Center.

Wong, J. K. K. (2004). Are the learning styles of Asian international students culturally or contextually based? *International Education Journal*, 154-166.

Yang, B. S. (1994). Morphosyntactic phenomena of Korean in role and reference grammar: psych-verb constructions, inflectional verb morphemes, complex sentences, and relative clauses. *Unpublished Doctoral Dissertation*. State University of New York at Buffalo.

Yu, G. (2009). Lexical diversity in writing and speaking task performances. *Applied Linguistics*, 236-259.

Yuan, F., & Ellis, R. (2003). The effects of pre-task planning and on-line planning on fluency, complexity and accuracy in L2 monologic oral production. *Applied Linguistics*, 1-27.

Zobl, H. (1984). Uniformity and source-language variation across developmental continua. In W. Rutherford, *Language Universals and Second Language Acquistion* (pp. 185-218). Amsterdam: John Benjamins.