ORAL FLUENCY DEVELOPMENT ACTIVITIES: A ONE-SEMESTER STUDY OF EFL STUDENTS

A Dissertation
Submitted
to the Temple University Graduate Board

In Partial Fulfillment
of the Requirements for the Degree of
Doctor of Education

By
Timothy Doe
December, 2017

Examining Committee Members

David Beglar, Advisory Chair, Teaching and Learning
Yoko Kozaki, External Member, Mukogawa Women's University
Steven Ross, External Member, University of Maryland
James Sick, External Member, New York University, SCPS

ABSTRACT

The development of speaking fluency is a major goal for many EFL language learners and several researchers have proposed frameworks for fluency instruction based on theories of cognitive science. It is unclear however, whether EFL students with restricted opportunities to use English outside of language classrooms can benefit from fluency development activities. This main purpose of this study was to determine whether EFL students' speaking fluency improved whilst participating in theoretically grounded fluency development activities. In addition, student use of formulaic language, participation in classroom activities, and repetition of previously used lexical items was examined in order to determine its relation to development in speaking fluency. 32 first-year Japanese university students from four intact discussion skills classes took part in the study, which was conducted over a period of 12 weeks. Data were collected by recording student performances in a variety of fluency development activities and interactive communicative tasks.

There were four main questions investigated in this study. The first research question was focused on long-term fluency development by using multi-level modeling to determine whether gains were made in fluency measures in four 2-minute speaking monologue tests that were conducted at regular periods throughout the semester. Three in-class performance variables - the amount of repeated words, the amount of tokens spoken, and the amount of formulaic language spoken, and three individual difference variables—willingness to communicate, extraversion, and first language fluency were also examined to determine if they were related to any growth observed. The second research question looked at short-term fluency development

across a time-pressured speaking activity and similarly used multi-level modeling with the same predictor variables. The third research question concerned the relationship of complexity and accuracy to the fluency measures derived from the monologue speaking tests. Finally, the fourth research question was an investigation of the relationship between the objective fluency measurements and subjective expert ratings.

The results indicated that the participants made very small, but significant gains on their mean length of pause across the monologue speaking tests. While none of the predictor variables had a strong relationship with this development, post-hoc analyses suggested that other fluency measures and oral proficiency level could have covaried with this growth. There was also a small but significant improvement in the phonation/time ratio, however, none of the predictor variables appeared to covary with this development. No other significant relationships were found in the long-term fluency measures. For the short-term fluency measures, significant growth was seen across the three deliveries of the speaking activity. While several of the predictor variables, most prominently repetition and the number of tokens spoken, had a significant relationship with this growth, a closer examination revealed that the degree of covariance was extremely slight.

The relationship among complexity, accuracy, and fluency also became more significantly correlated over time, suggesting that learners produced higher quality samples of language as the study progressed. Finally, the expert ratings had significant correlations with three of the five fluency measures, indicating that human raters were able to detect small differences in spoken fluency.

The findings of this study show that fluency can develop in instructed foreign language settings, however, the role of practice and repetition might be more complex than has been suggested in the research literature. This study provides some insight into that complexity and suggests a number of directions that can be followed to understand more about fluency development.

ACKNOWLEDGMENTS

This study could not have come about without the kind assistance of many people. First, I would like to thank Dr. David Beglar, my dissertation advisor, who provided me with continuous support throughout the duration of this project. Since entering the TUJ Master's program, Dr. Beglar has taught me a great deal about the field of second language acquisition, and several of the concepts that he introduced to me in his classes are taken up in this study. Dr. Beglar's enthusiasm for the field of second language acquisition is infectious, and after sharing my ideas with him, I always felt motivated to work harder and learn more. I truly appreciate the time he took out of his busy schedule to help me on many aspects of this dissertation.

Next, I would also like to thank Dr. Steven Ross, whose comprehensive knowledge of research methodology and statistical analyses helped me to find the best methods to analyze my messy data. I very much appreciate his generous assistance while I was analyzing the multi-level models, especially during a busy summer vacation. His kind encouragement helped me to stay positive and continue working, especially when I was learning about new techniques that I found challenging. I feel very grateful for the experience of taking his classes in the doctoral program, which greatly deepened my understanding of the different ways in which second language research can be conducted.

I would also like to thank the other members of my dissertation defense committee. Dr. Yoko Kozaki's published research was very helpful when analyzing the multi-level models in this study, and I very much appreciate her patient explanation of how to present these models in my results section. Dr. James Sick gave me some very important feedback on the statistical analyses and their possible interpretations, especially with the Rasch model analyses. I feel very lucky to have had such a supportive and knowledgeable committee, and I learned a great deal from their thorough and detailed feedback.

I would also like to thank the members of my doctoral cohort, especially Todd Leroux, Rick Romanko, and Satoko Suzuki. Over the past few years, their friendship and support has energized me to continue working, and I am continually inspired by their dedication to research and presenting at academic conferences. I feel very lucky to have been able to discuss my ideas with them, and their feedback always helps me to refine my thinking. I would also like to thank my colleague Chie Ogawa, who introduced me to several fluency articles that were very influential in this study, and who always encouraged me to complete the study.

I would also like to thank my colleagues who helped me a great deal with the implementation of this study. First, I would like to thank Stephen Hurling, who did much of the work in establishing the discussion program in which this study was carried out. Without Stephen's help and expertise, I would not have been able to complete this study. I would also like to thank Matthew Livingston, who taught me a great deal about communicative language teaching and program management. Finally, I would like to thank my colleagues at Kwansei Gakuin University, Ian Handsley, Joshua Jodoin, Stewart Markel, Stephen McNamara, Stephen Ryan, and Andrea Tan, who kindly offered to act as expert raters, and took time out of their busy schedules to help me with this project.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

In this chapter I first outline the background to this study and then describe the main issue which I seek to investigate: whether students in instructed English as a Foreign Language (EFL) settings can develop their speaking fluency with limited opportunities to use English outside of the classroom. I begin by explaining why learning more about how students develop speaking fluency in EFL contexts is an important issue to address, briefly outlining current research into fluency development, especially concerning the provision of opportunities for students to repeat language that has been produced on prior occasions. Following this, the specific aims of this research are outlined. A description of the potential significance of the findings of the study is followed by a brief discussion of the intended audience and delimitations.

The Background of the Issue

Developing the ability to speak effectively in a foreign language is one of the main goals of many learners. To do so, students must develop some degree of speaking fluency; that is, they must be able to produce language smoothly and relatively quickly. The term fluency can be used to refer to many aspects of spoken language, including the speed, quality, content, and appropriacy of an utterance (Fillmore, 2003). In second language acquisition (SLA) research, three common distinctions are made among utterance fluency, referring to pausing and hesitations, speed fluency, referring to how much language a speaker can produce in a given time, and repair fluency, referring to repetitions and self-corrections (Tavakoli & Skehan, 2005). Developing these aspects of speaking fluency is crucial if learners are expected to be able to communicate in real time; to do so, they must be able to organize and produce language quickly and smoothly.

Many SLA theorists believe that speaking fluency is the result of automatized language production skills (Kormos, 2006; Segalowitz, 2010). As the term implies, automaticity refers to behavior that is fast and able to be carried out simultaneously with other tasks. Adult speakers using an L1 can usually produce speech while doing everyday tasks such as driving and cooking. The situation for most L2 speakers however, is very different. Even if speakers concentrate a great deal on their speech production, the L2 is usually produced much more slowly and requires a great deal of effort. One possible explanation for this comes from skill acquisition theory, which states that skills come about from large amounts of practice (DeKeyser, 2007a). Following this idea, L2 speakers will not be able to produce language automatically until they have accumulated many experiences using the target language. Given the large amount of time needed to do so, it is not surprising that many adult learners fail to gain high levels of speaking fluency in foreign languages.

The implications of the above theories are that foreign language educators need to provide opportunities for learners to produce language in order for them to develop speaking fluency, especially in settings where learners do not have opportunities to use the L2 outside of the language classroom (Derwing & Munro, 2013). This need is likely to be more pronounced in settings where many learners have limited exposure and access to the L2. In recent times, communicative language teaching approaches such as task-based language teaching (TBLT) have become more widespread in EFL settings (Shehadeh, 2012) because TBLT presents a framework in which teachers can organize activities for learners to produce language in meaningful situations, carrying out tasks similar to those that they might use in the outside world. Such activities can potentially provide learners with the opportunity to practice the L2 and develop their speaking skills.

It has been important for researchers to determine therefore, which kinds of tasks are the most effective for developing specific areas of language abilities, such as speaking fluency. The complexity, accuracy, and fluency (CAF) framework has been used to investigate how different variables can affect the quality of language students produce when they carry out communicative activities (Housen & Kuiken, 2009). This framework divides language into three main categories: the complexity of the structures, vocabulary and syntax used by the speaker, the accuracy of speech in comparison with native-speaker norms, and the fluency with which the speech is produced. This line of research is promising for language teaching pedagogy because if more is understood about which kinds of activities can affect language development, syllabi could be produced that can be specifically tailored to meet learners' needs.

TBLT approaches have been the main kind of activities used in CAF studies. The variables that have been studied can be divided into three main categories: task complexity, for example, the amount of planning time given to students before undertaking the task, task conditions, for example, whether or not learners are familiar with other students they complete the task with, and task difficulty, for example, learners' levels of proficiency in relation to the task they are being asked to do

(Robinson, 2001). Findings from many of these studies have suggested that communicative activities can be manipulated in order to produce or develop specific language skills. This line of research has been promising for SLA researchers interested in learning more about how to help students develop their speaking fluency.

While TBLT provides a good framework for studying aspects of CAF, there has also been specific research into communicative activities designed to focus on the individual components of the CAF framework. In terms of activities that promote speaking fluency, research on the 4/3/2 speaking technique (Maurice, 1983) has shown that opportunities for repeating speaking tasks potentially influence learners' speaking fluency as well as accuracy and complexity (Nation, 1991). This task involves a speaker giving a speech three times to three different listeners. On each occasion, the time available for speaking decreases by one minute, however, the speaker attempts to convey the same information. It has been argued that through this provision of opportunities to repeat the same information, learners can practice and become more fluent in using specific language items and structures (Maurice, 1983). This claim has been supported by the findings of a recent study that reported long-term fluency gains for speakers in the experimental condition who were repeating the same information in their 4/3/2 speeches, in contrast with a lack of gains for speakers in the control group who were asked to change topic each time they spoke (De Jong & Perfetti, 2011). Such findings suggest that activities that encouraging or requiring learners to repeat vocabulary could be an effective means of developing their speaking fluency.

There have also been teaching methodologies proposed that aim to develop the speaking fluency of learners. Gatbonton and Segalowitz's ACCESS model (2005) consists of three phases of sequenced activities that have the specific purpose of building automatic fluency. First, the creative automatization phase comprises of tasks that enable the teacher to elicit target formulaic language from students and allow them to use such utterances repeatedly. Second, in the language consolidation phase, learners complete fluency and accuracy tasks to focus on areas of difficulty they encountered during the first phase. Finally, in the free communication phase, students complete communicative activities that allow them to produce the target phrases in similar but differing contexts. While this model has not yet been empirically tested, it follows one of the main tenets of skill acquisition theory in that repeated opportunities to practice target skills are necessary in order to build fluency. This model is attractive in that it expands upon the 4/3/2 technique and builds in opportunities for repetition across a variety of activities within a class. In addition, this model incorporates the direct teaching of formulaic language, the use of which has been argued to facilitate fluency development (Wood, 2010).

While it could be possible to manipulate activities to develop different aspects of students' speaking abilities, one factor that must be considered is the learners themselves. CAF research has shown that learners respond to different tasks in different ways, and when performing tasks that have been designed with specific intentions, "the influence of learners on the task can jeopardize the task designer's goals" (Murphy, 2003, p. 358). Many studies involving task design and CAF involve a small number of tasks that have been theorized to assist learners in producing more accurate, complex, or fluent language. However, learners can choose to focus on different aspects of their language production, and it could be the case that certain types of learners might be resistant to task design. In the case of fluency development, some learners might choose to focus more on accuracy and complexity, and thus make less progress than their peers who are more concerned with developing speaking fluency. Furthermore, because many studies of fluency development have shown that larger gains are often made by study abroad students, particularly those with a strong knowledge of grammatical forms prior to their sojourn (DeKeyser, 2010), it remains unclear whether speaking fluency can significantly develop in EFL settings, where learners have less access to the target language and spend considerably less time using it for communication.

Statement of the Problem

It is well accepted that students must engage in large amounts of meaningful practice using the target language to develop speaking fluency. Researchers who have studied English as a second language (ESL) students who failed to develop speaking fluency in target language settings have often attributed students' lack of success to the paucity of exposure they had to English outside of their language classes (Rossiter, Derwing, Manimtim, & Thomson 2010). In order to enable students to become more fluent, specific teaching approaches and techniques have been developed for use in language classes. One condition common in many of these approaches is the need to create activities that are inherently repetitive (Segalowitz, 2010, p.176). This conception of repetition differs from rote memorization and drills common to teaching methodologies such as the Audiolingual approach. Activities that are inherently repetitive give students the opportunity to repeat previously used items and forms while engaged in meaningful interactions. Results from research into fluency activities and task repetition have suggested that provision of such opportunities might be beneficial to students (Bygate, 1991; De Jong & Perfetti, 2011). However, it must be emphasized that these studies have been of higher-proficiency students in ESL settings who have had many opportunities to practice and refine their speaking skills outside of the language classroom. To date, there has been very little empirical investigation of whether meaningful repetition can benefit EFL students who have lower levels of speaking proficiency and fewer opportunities to use the target language.

Learning more about the development of students' speaking fluency in EFL settings is crucial if courses are to meet the aim of providing students with the ability to use English meaningfully in real world settings. Several recent researchers have examined student performance in different kinds of TBLT conditions, analyzing how factors such as preparation time and provision of task repetition might affect student complexity, accuracy, and fluency (CAF). The results of such research commonly suggest that certain conditions or task types are beneficial for developing specific aspects of CAF, yet one area that has been relatively neglected is the performance of students across several tasks and communicative activities. Furthermore, while it has been theorized that the provision of formulaic utterances, repetition, and extended opportunities to practice communication aids fluency development, it remains to be seen the degree to which individual students can affect this outcome, and whether those who take the opportunity to speak more and repeat more of the same items make greater gains than their peers who do not.

Purposes and Significance of the Study

This study is designed to determine the degree of variability with which Japanese EFL students' fluency develops over an academic semester. Several studies have shown that both learners in instructed settings and those who live abroad have not made gains in speaking fluency, in some cases over periods of several years. The findings of these studies were based on data collected at quite large intervals of up to one year. In contrast, in this study I investigate fluency development with a greater density of data collection points in order to discover the degree to which EFL students' speaking fluency varies over the course of an academic semester.

Performance variance is important to determine for two practical reasons. First, many EFL students take classes to become more fluent speakers; however, whether such classes can have short-term effects on speaking skills is unclear. The results of this study will give insight into the issue of how much time and attention EFL students need to pay to developing greater speaking fluency. Second, many EFL courses have adopted communicative approaches in order to develop students' speaking abilities, but little is known about the effectiveness of such pedagogical approaches. It is important to learn more about how suitable communicative activities are for the development of speaking abilities in an EFL setting.

The second purpose of this study is to investigate which factors of student performance are related to gains in fluency, specifically repetition, quantity of speech, and use of formulaic language. Theories of automaticity and speaking fluency claim that large amounts of practice and repetition are necessary to develop automatic speaking skills (DeKeyser, 2007). In addition, several prominent researchers of speaking fluency (Gatbonton & Segalowitz, 1988, 2005; Wood, 2001) have argued that these factors are instrumental in fluency development. To date, however, there has been little research into the potential influence these factors might have on fluency development. In addition to performance factors, individual difference factors that might be associated with L2 fluency, such as extraversion, willingness to communicate, and L1 fluency, are investigated, as there has been little research regarding how individual differences affect speaking fluency (Segalowitz, 2010).

These findings can provide both theoretical and practical benefits. From a theoretical perspective, understanding how repetition, quantity of speech, and use of formulaic language affects the development of speaking fluency would provide evidence to support several teaching methodologies that have been designed to incorporate findings from research into automaticity and speech production. From a practical perspective, learning more about whether students' levels of spoken fluency can benefit from attention to certain aspects of performance will be of use to curriculum designers and teachers who seek to provide effective fluency development courses. Furthermore, as this study involves student performance across several activities within a lesson, the results will be more detailed than previous studies that have been focused on either one fluency activity or several unrelated tasks.

A further purpose of this study is to examine student gains using the CAF framework by investigating whether students who make fluency gains also develop greater accuracy and complexity, in addition to investigating whether students who do not improve their speaking fluency might do so in other areas of the CAF framework.

From a theoretical perspective, these findings add to the ongoing research into how different aspects of CAF are affected by both task design and contributions from the learners themselves. From a practical perspective, the findings might reveal some of the characteristics of students who are more inclined to develop speaking complexity, accuracy, and/or fluency, and therefore be of use to teachers who need to identify student needs in relation to learning goals and objectives.

The final purpose of this study is to investigate the degree to which expert ratings of perceived fluency correlate with temporal measures of utterance fluency. Several researchers have investigated this question (Derwing, Munro, Thomson, & Rossiter, 2009; Lennon, 1990; however, the data used in these studies has been collected at fairly wide intervals. In this study I attempt to determine which aspects of spoken fluency are noticed most by expert raters when listening to speech samples that were collected at relatively short intervals.

The potential benefits of such findings are to identify which factors of spoken language are most salient to expert raters when attending to matters of fluency. Such findings can contribute to the ongoing investigation into which factors of spoken fluency are most salient to perceptions of spoken fluency.

Theoretical Perspective

While this study is primarily concerned with the development of speaking fluency, it is necessary to place fluency within other aspects of language performance in order to determine the exact nature and degree of variation of student development. The conceptual model of complexity, accuracy, and fluency provides a framework of spoken performance that is able to account for many different aspects of speech. Many SLA researchers believe language proficiency to be "multicomponential in nature, and that its principal components can be captured by the notions of complexity, accuracy, and fluency" (Housen, Kuiken, & Vedder, 2012, p. 1). Although the CAF construct has not been uncontroversial, this framework has been central to many studies of second language performance and development since the 1990s (Housen & Kuiken, 2009, p. 462). While researchers continue to focus on attempting to identify the most valid and reliable measures of CAF, the current state of knowledge provides a strong basis to describe and measure aspects of language performances, and has been argued to maximize the variance among language speakers or writers (Skehan, 1998).

One of the biggest challenges for researchers using the CAF framework has been to define each of the three concepts. Accuracy, has largely been seen as "the degree of deviancy from a particular norm" (Housen & Kuiken, 2009, p. 463), and involves quantifying and classifying errors made by a language user. The term can also be expanded to include the appropriacy and acceptability in terms of standard language norms. Complexity has been generally described as "the extent to which learners produce elaborated language" (Ellis & Barkhuizen, 2005, p. 139). Bulte and Housen (2000) separated L2 complexity into three sub-components, propositional complexity, concerned with the number of idea units encoded into language; discourse-interactional complexity, concerned with the type and number of turns learners take during interactions; and linguistic complexity, concerned with structure and syntax. Finally, fluency is concerned with the temporal features of speech, especially "the flow and smoothness of delivery" (De Jong & Perfetti, 2011, p. 534).

Under this framework, fluency is often broken into two subcomponents, temporal features of speech, such as the rate of speech, and hesitation phenomena produced during speech, such as false starts (Ellis & Barkhuizen, 2005, p. 156).

Perhaps one of the reasons that CAF research has become more prominent in SLA research is that although it is based on an information processing view of cognition, it has been used by researchers from a wide range of backgrounds, from Universal Grammar (Towell, 2009), to Dynamic Systems Theory (Larsen-Freeman, 2006). One of the benefits of this approach is that is can be utilized by researchers of different theoretical orientations. A second benefit is that if CAF features are examined at the same time, they can capture elements of language performance that would have been missed if only a single element was examined. For example, if a researcher was solely concerned with the correct use of a particular language form, developments in other areas of the framework might be missed. CAF measures can also be used to identify if learners have different goals when they are using language, as some learners might concentrate solely on meaning, resulting in an inaccurate but fluent performance, whilst others might be concerned with using correct grammar forms, and thus produce less fluent but highly accurate speech samples. For this reason, it is important to look at a variety of CAF variables in order to gain a more comprehensive picture of student performance, although as noted by Ellis and Barkhuizen, the wide range of measures that have been used in CAF studies make it difficult to generalize findings (2005, p. 163).

The Audience for the Study

The potential theoretical findings of this study should be of interest to second language acquisition researchers, curriculum designers, and practicing teachers. For the first group, the investigation of the relationship between repetition and fluency addresses current theories of speaking development. The majority of studies into repetition and speaking have been either cross-sectional or with a small number of data collection sessions. This study expands upon these findings by providing information about how students develop speaking fluency across weekly classes over an academic semester, in addition being one of the few studies that examines how interaction in communicative tasks could affect fluency development.

The practical benefits of this study will be of interest to school administrators, curriculum designers, and teachers who seek to develop and maintain high levels of student speaking fluency. Speaking fluency has often been a neglected element of language classes and little is known about how to meet the needs of EFL students who seek to develop speaking fluency. Following student development across a variety of activities conducted over a semester will offer insights into the possible advantages and disadvantages of conducting fluency activities in class.

Delimitations

The participants' first language, Japanese, and their proficiency levels are factors that affect the generalizability of this study and need to be highlighted as delimitations of the study. There is a large degree of language distance between Japanese and English, as the two languages have very different syntactic structures, and Japanese is a mora-timed language whereas English is stress-timed. Therefore, different results might be found for students from different L1 backgrounds, particularly those whose L1 is genetically related to English. Furthermore, the majority of students in the study were at a similar level of language proficiency. Automatic processing of language, which is related to fluency development, is theorized to follow a power law of practice where initially larger gains gradually taper off over time (Logan, 1992). The results of this study, therefore, cannot be generalized to groups of different proficiency levels because their rates of development might be very different to those of the students who took part in this study.

Because this study took place in a very specific environment, the second delimitation that needs to be mentioned is the sample group, which is made up of first-year students taking a compulsory English discussion class at a private Japanese university. Furthermore, these students are of similar levels of language proficiency and the large majority comes from the same cultural background. These students were also placed in small seven to nine-member discussion skills classes that required a high level of student-to-student interaction, approximately two-thirds of a 90-minute lesson. It is possible that different results would be found with Japanese students in different educational settings or language learners who come from different cultural backgrounds and/or with different levels of language proficiency.

Another limitation of this research is that of the complexity, accuracy, and fluency variables that were selected for analysis. Given that there is a wide range of potential variables for each of the three constructs, decisions had to be made about which might be the most appropriate means to determine changes in student behavior.

It is possible that if different variables had been selected, the results might differ from those described in this study.

Organization of the Study

Chapter 2, Review of the Literature, is made up of six main sections: Speech Production, Automaticity, Fluency, the CAF framework, Extraversion, and Willingness to Communicate. At the end of the chapter the gaps in the literature, purposes of the study, and research questions are presented. In Chapter 3, Methods, I describe the participants, instrumentation, procedures, and analyses used in the study. Chapter 4, Preliminary Analyses describes how the data were screened and prepared before the main analyses. In Chapter 5 Results, I detail the results of the findings, while in Chapter 6, Discussion, I interpret the findings. Finally, in Chapter 7, Conclusion, I summarize the main findings, discuss the limitations of the present study, offer suggestions for future research, and make concluding comments.

CHAPTER 2

REVIEW OF THE LITERATURE

In this chapter I review the literature on speech production, automaticity, formulaic language, fluency, and research into the CAF framework. At the end of the chapter I identify the gaps in the literature, state the purposes of this study, and list the research questions this study is designed to answer.

Speech Production

In both L1 and L2 speech production research, the most influential model is Levelt's modular Speaking model, which describes how intentions are transformed into speech for adult monolingual speakers (De Bot, 1992; Kormos, 2006). It is a 'steady state' model, and therefore, does not account for language development, or for neurological disorders that might affect the ability of an individual to produce speech (Levelt, 1989). While the model is not comprehensive and has been revised several times, many researchers argue that it is the most complete and widely accepted model of speech production, as to some degree, it represents "the consensus view of the linguistic, psycholinguistic, and cognitive issues underlying the act of speaking" (Segalowitz, 2010, p. 8).

Levelt sees speech production as being the result of several components working in a modular fashion, similar to a computer algorithm processing information (see Figure 1). These components are seen as relatively autonomous and specialize in the functions they carry out. Each component feeds its output forward into the next in a sequential order. There are two major kinds of components; the knowledge components, which are the circular ellipses in Figure 1, and the production components, which are the rectangular boxes in Figure 1. The knowledge components represent declarative knowledge, including a speaker's knowledge of the world, as well as models of discourse and interactional situations. A separate language-specific knowledge component, the lexicon, is included in the model and is used at later stages of language production and monitoring. In contrast, the production components contain procedural knowledge that process the information drawn from the knowledge stores to produce speech.

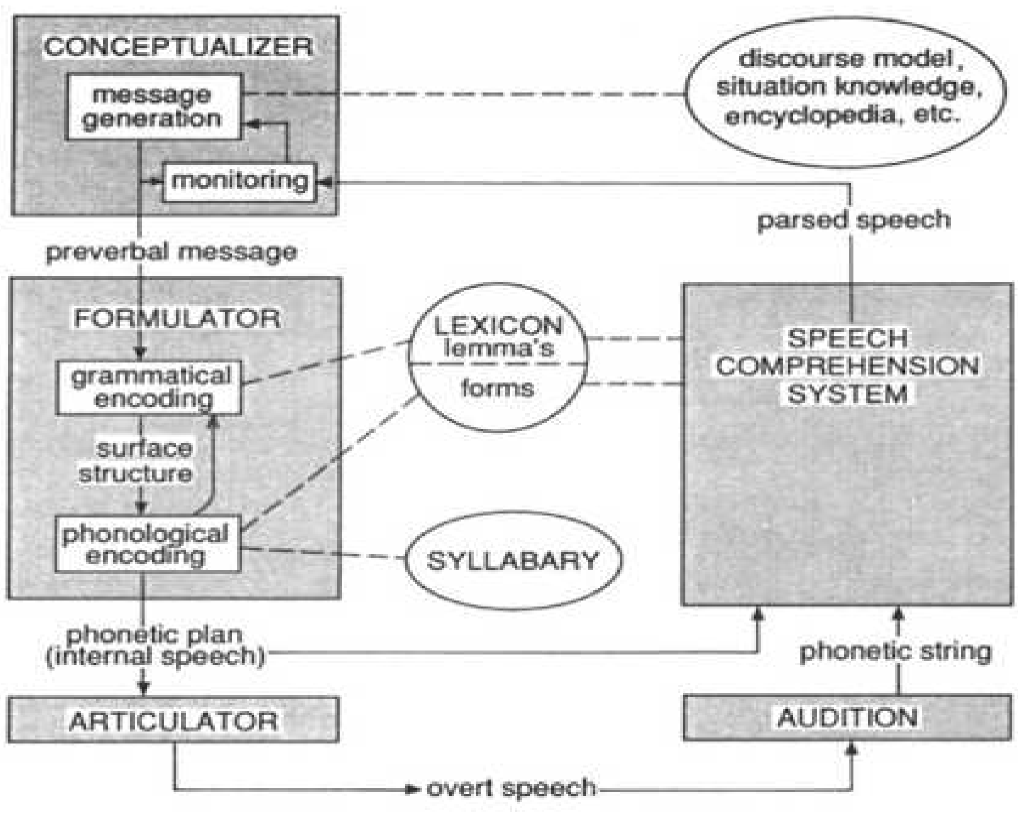

Figure 1. Levelt's blueprint for the speaker (from Levelt, 1989, p. 9).

This initial process of transforming a conceptualized utterance into speech involves converting a preverbal message into linguistic material, which involves sequential processes of encoding. As implied by its name, the preverbal message contains no linguistic information. This message is moved into the formulator and transformed into language by applying lexical, grammatical, and phonological rules in serial order. The appropriate vocabulary is selected and grammatical and phonological rules are applied to create a speech plan, which is forwarded to the next component, the articulator. The articulator converts the speech plan into spoken language which is attained through motor control of the necessary muscles needed to produce sound. In addition, the speech plan forwarded from the formulator is also fed into the speech-comprehension component, which checks for any errors and can give feedback into the conceptualizer if corrections or adjustments are necessary. Later versions of Levelt's model include two further forms of monitoring; monitoring of the preverbal plan with the intended message, which takes place in the conceptualizer, and pre-articulatory monitoring before speaking to check encoding errors with vocabulary choice, which takes place in the formulator.

Each component in Levelt's model has a specialized function, and operates in a relatively autonomous manner at very high speeds, which Levelt describes as an "automatic, reflex-like way" (1989, p. 2). Furthermore, the system can use parallel processing when producing speech. For example, as soon as the conceptualizer has sent the pre-verbal message to the formulator, it can immediately begin work on the next information that will be produced, while the formulator and the other components are simultaneously processing information. Such a system can explain the ability of humans to produce speech smoothly and rapidly.

Levelt's model, however, is not without controversy, especially in relation to the serial manner in which its components operate. Kormos (2006) summarized competing arguments from alternate theories of speaking and concluded that although Levelt's model seems the most valid, it seems likely that not all information is processed in a one-way serial manner. It has also been pointed out that in neurobiological terms, a strictly modular approach to language processing seems unlikely, given the two-way interaction found among neuronal networks in many studies (Pulvermuller, 2002). Furthermore, more recent views on language influenced by dynamic systems theory argue that declarative knowledge does not exist, meaning that all knowledge is procedural. This claim is based on the view that all knowledge is composed of processes, rather than the storage metaphors employed in Levelt's model (Lowie & Verspooer, 2012).

While Levelt's model is not unproblematic, it still represents the most comprehensive model of speech production and many of its claims have been supported by empirical research (Kormos, 2006; Pulvermuller, 2002). SLA researchers, therefore, have been interested in how the model might apply to multi- and bilingual speakers, and have theorized how the model might be adapted to account for knowledge of more than one language. De Bot (1992) has argued that with minor modifications, Levelt's model can account for features such as code-switching and cross linguistic influences that are common to bilingual speech, and created an adapted model of a bilingual speaker.

Furthermore, in an analysis of De Bot's adaptation of Levelt's model, Segalowitz (2010) identified seven fluency vulnerability points that could impact the ability of second language speakers to produce speech. These are:

- not knowing the lexical items that are necessary in microplanning
- not having adequate linguistic knowledge when creating a surface structure when grammatically encoding the preverbal message
- being unable to retrieve lexical items
- not having adequate morpho-phonological knowledge when creating the phonological plan
- not being able to automatically access the gestural scores - that is, metrical rules and intonation contours - associated with the second language
- not being able to automatically produce the gestural scores associated with the second language
- having to devote more of their attentional resources for self-monitoring compared with L1 use

While some of the above factors are related to a lack of knowledge of certain aspects of the L2 and cross-linguistic transfer between the L1 and L2, many of them relate to automaticity, which refers to the efficiency with which the mind can process language. In order to produce speech fluently, speakers must be able to automatically process and produce language.

Skehan (2009) has also argued that Levelt's model can be a useful means of conceptualizing some aspects of second language speech, in particular, the distinction between the conceptual and linguistic processes involved in speech production.

Skehan argued that issues related to task complexity can affect processing in the conceptualizer, whilst language-related issues affect the functioning of the formulator. Skehan has argued that this dichotomy is important to take into account when considering task difficulty, as different features of language tasks will present different kinds of challenges to second language learners.

To summarize, while Levelt's speaking model does not comprehensively account for all of the factors related to second language speech production, it has been widely accepted as the most comprehensive model of L1 speech production. Furthermore, many SLA researchers have argued that the model can also account for some of the differences between L1 and L2 speech. In particular, having incomplete stores of lexico-grammatical knowledge of the L2 can cause problems for L2 speakers to produce fluent speech. Therefore, it is important for SLA researchers to learn more about how speech production in the L2 can become automatized.

Automaticity

One feature common to all adult speakers using their L1 is that they can produce speech at an extremely fast rate, making it appear that speech is produced automatically. In Levelt's model, apart from message generating and monitoring, all speech production processes "are assumed to be automatic to a large degree" (Levelt, 1989, p. 28). This could account not only for the speed and relative ease with which speakers can use their L1, but also the difficulty that many speakers have explaining the grammatical rules that govern the L1. In terms of speaking an L2, however, most speakers speak slowly and need to pay attention to what they are saying, often applying grammatical rules or consciously recalling vocabulary items from memory. Research findings have suggested that even advanced L2 speakers cannot produce speech in their second language automatically (Kormos, 2006).

Automaticity is often assumed as simply being fast processing; however, there are several other properties that can be used to describe automatic processes (Segalowitz, 2003). Segalowitz noted that these include: ballistic, meaning that even with conscious effect, an individual cannot help processing the information; load independent, meaning that processes operate regardless of the amount of information that needs to be processed; effortless, meaning that information can be processed without attention, leaving the speaker free to engage in other tasks; and unconscious, meaning that speakers are unable to recall specific previous encounters with the information. Research has shown that all of the above characteristics, save load independence, suggest automatic processing (Kormos, 2006, p. 40). Segalowitz, therefore, urged researchers to avoid reducing fast processes to automatic ones, because as well as an increase in speed, automatized processes "imply greater processing *efficiency*" (2010, p. 79, italics in the original).

There have been several attempts to build a theory describing the mechanisms that allow for automaticity to develop. These theories can be classified into two major groups: rule- and instance-based theories (Kormos, 2006).

Rule-Based Theories of Automaticity

Anderson's ACT (adaptive control of thought) theories are examples of a rule-based theory. These theories hypothesize that automatic behavior results from declarative knowledge becoming proceduralized before finally becoming automatized (Anderson, 1992; Anderson & Lebiere, 1998). Declarative knowledge is acquired through observation, reading, and discussion. At this stage, the learner has not experienced trying to perform the skill. The next stage, procedural knowledge, is when the declarative knowledge is acted upon, and the learner attempts to perform the target behavior. This transformation is known as proceduralization, where *knowledge that* becomes *knowledge how* (DeKeyser, 2007b). During the process of proceduralization, it is thought that the individual components of the skill are "chunked" into larger memory representations. The change, therefore, is not simply a quantitative speeding up of language processing, but a qualitative restructuring of knowledge. Chunked language representations reduce the demands on working memory because instead of having to recall each individual step, the learner can access a process of linked steps.

The final process is automatization, where the learner becomes able to perform the task fluently, unconsciously, and without error. Kormos (2006) states that the automatized result of proceduralized knowledge might explain why speakers may no longer be able to retrieve the declarative knowledge that was used to develop the skill, giving the example that advanced level English language learners often forget grammatical rules they memorized earlier in the language learning process, yet are still able to use them with a high level of accuracy. This stage is hypothesized to be the result of a significant amount of practice, however, learners can still err when performing skills that have been fully automatized (Anderson, 1992).

A salient weakness, however, in rule-based theories, is that it assumes all learning is the result of declarative knowledge being proceduralized. Kormos (2006) argues that there are probably instances where a memorized chunk may replace production rules in memory. She also cites computer simulation studies of language acquisition, which have suggested a different developmental process. In these studies, associative learning programs have been given a large amount of exemplar chunks of language. Through the processing of these exemplars, the associative learning programs have been able to abstract the grammatical rules underlying them. These findings highlight that besides the proceduralization of declarative knowledge, there may be other processes involved in the automatization of language representations.

Instance-Based Theories of Automaticity

In contrast to rule-based theories, advocates of instance-based theories have argued against the distinction between declarative and procedural knowledge. Logan's instance theory (1988) hypothesizes that novices use algorithms when completing tasks with which they have little experience. In the case of L2 learning, this would be the case of a learner applying grammatical rules to construct an utterance. However, once the learner begins to accrue experience, the same tasks might be able to be performed by memory retrieval of the solution rather than by referral to the algorithm. In contrast to rule-based theories in which automaticity is the result of the conversion of knowledge, instance-based theories posit that fluent speakers retrieve language representations from memory, and as these memory-based processes become stronger, they replace knowledge of rules and algorithms. Support for these theories can be seen in arguments of formulaic language, such as Pawley and Syder (1983), which argued that despite the large amount of options language users have, most speech is formulaic and predictable.

Similar to rule-based theories, however, it appears that instance-based theories of automaticity also cannot offer a comprehensive account of speech production. Kormos (2006) points out that instance-based theories cannot account for learners' abilities to generalize learned skills to unique contexts, that is, people are still able to produce language in novel settings for which they may have little or no memory representations. She also points out that instance-based theories have little to offer in terms of memories of language representations are stored and retrieved.

Summary

Due to problems with both explanations of the nature of automaticity, Kormos (2006) has argued that an explanation of this phenomenon must be able to account for both rule- and instance-based processing. In terms of lexical retrieval, she cites connectionist theories of knowledge retrieval and argues that the memory based processing outline in instance-based theories of automaticity are more convincing, given that visual stimuli can activate memories of specific words or chunks. However, in terms of syntactic and phonological knowledge, Kormos cites research that has shown a process of proceduralization in L1 acquisition that shows learners automatizing syntactic rules. While no research has been conducted with phonological knowledge, she argues that a similar process may occur. Therefore, it may be the case that both rule- and instance-based processing is used when learners automatize their language production skills.

While the actual nature of automaticity is still in dispute, few would argue that developing automaticity with language is an important need or goal for many speakers. Many L2 researchers have referenced Anderson (1992) and Logan's (1988) theories in addition to other practice theories to explain the process of how knowledge becomes automatized (DeKeyser, 2007b; Schmidt, 1992). In accordance with the above theories, L2 automaticity is thought to develop from extensive practice, although this aspect of acquisition has been largely overlooked in most of the SLA literature (DeKeyser, 2007a). Practice, like fluency is a term that has multiple meanings, however, skill acquisition theorists believe that effective practice is that which focuses on improving skills that are already known. DeKeyser (1997) studied the acquisition of L2 grammar, showing that learners follow the power law of learning, where reaction time and error rates follow a mathematical function, beginning with rapid improvement before slowing down and stabilizing. He argued that this process shows learners moving from declarative to procedural knowledge followed by a slower transition to automatization. While there has been little research in this area, these findings support the popular notion that to become fluent speakers, students need large amounts of practice with the language items that they have already been taught (Nation, 2001).

To summarize, automaticity in speech production is not simply producing speech quickly, it is the accumulation of a speaker's experiences using the language, which results in qualitatively different speech than less experienced speakers. This can often be seen when comparing the speech of a speaker's L1 and L2. While researchers have been able to describe the features of automatic speech, specifically being ballistic, effortless, and unconscious, there is still a lack of consensus as to how speech production becomes automatized in the human brain. The two leading theories are rule-based (Anderson, 1992), and instance-based (Logan, 1988), although both theories have problematic aspects. Despite these shortcomings, there is much theoretical support that suggests L2 speakers need large amounts of practice to develop the degree of automaticity necessary to become a fluent speaker of an L2.

Formulaic Language

One result of extensive practice with language use is that items that are frequently used together, such as the two words that make up the phrase *excuse me* become chunked into a single item. It has been argued that the ability of native speakers to produce speech at fluent rates is a result of being able to access chunks of stored language rather than having to generate each sentence using their knowledge of grammatical and syntactical rules (Pawley & Syder, 1983). The ability to chunk language frees up limited attentional resources allows speakers to devote their attention to *what* they say rather than *how* to say it. This can be seen in Logan's instance theory, where "lexical items that often occur together tend to form chunks (higher order phrases or clauses), and when the conceptual specifications call on them, they are retrieved as one unit" (Kormos, 2006, p. 46). These chunks are also known as formulaic language (Wood, 2010; Wray, 2002).

While there is has been a considerable amount of difficulty in precisely defining formulaic language (Wray, 2002), one common feature noted by many researchers is that the sequence of words is retrieved from memory as a single unit (Kormos, 2006; Richards & Schmidt, 2002; Wray 2002, 2008). While formulaic sequences and chunks have been studied in a variety of L2 settings, little has been researched in regards to their relationship with speaking fluency. Wood (2010) recorded 11 ESL students completing a narrative retelling task outside of their classrooms over six months. The results suggested that the use of formulaic language increased as learners became more fluent speakers. While Wood's study is probably the most comprehensive to date, other studies of ESL students have also partially attributed fluency development to an increase in the use of formulaic language (Towell, Hawkins, & Bazergui, 1996). To date, however, there have been no major empirical investigations of whether the direct instruction of formulaic sequences can enhance fluency instruction.

In summary, the ability of speakers to chunk units of language together has been argued to be an important factor in the ability to speak fluently. It is theorized that highly proficient speakers have access to a large amount of formulaic language that allows them to pay attention to other features of the context in which they are communicating. However, despite the importance of formulaic language for producing fluent speech, there has been little research into how speakers of L2s use and develop chunked language forms. Initial results of research looking at L2 speech production and chunking suggest that the use of formulaic language correlates with increases in fluency, although further research needs to be conducted in this area.

Fluency

Studies in SLA that have investigated the temporal features of the spoken language with which L2 speakers can produce speech have used the term *fluency* to describe the object of their focus. However, researchers have not been able to arrive at a single definition for fluency and no reliable indicators of fluency have yet been found or agreed upon (Segalowitz, 2010, p. 46). Historically, some of the central concepts connected with fluency have been *flowing*, *smoothness*, *rapidity*, and *automaticity*, (Koponen & Riggenbach, 2000, p. 6). Further definitions have referred to spoken fluency as spontaneous language, natural connected speech, the ability to manipulate language automatically, and being able to produce speech that is socially or pragmatically appropriate (Koponen & Riggenbach, 2000, p. 8). Lennon, (1990) suggests distinguishing between fluency in a 'broad' sense when referring to features of fluency that relate to overall language proficiency, and fluency in a 'narrow' sense when referring to one specific factor of a speaker's output, related to the flow of their speech. Components of narrow fluency are often contrasted with those of accuracy and complexity in studies that use a CAF framework.

However, for researchers studying speaking fluency, the narrow sense of fluency needs to be broken down into more specific groups. Segalowitz (2010) suggested the following labels to avoid confusion: cognitive fluency, utterance fluency, and perceived fluency. Cognitive fluency refers to "the speaker's ability to efficiently mobilize and integrate the underlying cognitive processes responsible for producing utterances with the characteristics that they have" (p. 48). In Levelt's model, this term would be used to describe the speed with which each component can pass on information to the next. Due to the difficulty in measuring cognitive fluency, few researchers have investigated it in relation to L2 speaking. However, researchers have used computer software to measure reaction times to written prompts as a means to investigate how efficiently individuals can process language as a measure of cognitive fluency (Segalowitz & Freed, 2004).

Utterance fluency refers to the temporal qualities and characteristics of speech performances that can be objectively measured, for example, by counting how many syllables a speaker can produce in a minute, or measuring the length of time taken for pauses while producing speech. Many of the measures used in SLA research investigating utterance fluency were taken from first language researchers such as Goldman Eisler (1968), who were interested in the degree of variation that native speakers of English exhibited when they were speaking. Utterance fluency has been divided into three further categories: breakdown fluency, relating to pauses and silences; speed related factors, such as speech rate and mean length of run, which is typically the mean number of syllables between pauses; and repair fluency, which encompasses features such as repetitions, reformulations, and self-corrections (Tavakoli & Skehan, 2005). The most common utterance fluency features that have been used in research are shown in Table 1.

Finally, perceived fluency refers to listeners' perceptions of the speaker's level of fluency. In some speaking tests, for example, a panel of judges rate the fluency of the speaker based on a rubric, resulting in an overall impression of speaking ability rather than a quantitative analysis of speech. A series of fluency studies have had Derwing & Munro, 2013). However, while raters can be trained to reliably assess utterance fluency, research has shown that judges can place differing levels of significance to factors such as speech rate and lexical choice (Kormos & Denes, 2004).

Table 1. *Overview of Fluency Measures (from Segalowitz, 2010, p. 6)*

Measure (units)	Definition
Speech rate (syllables/minute)	60 sec./min. times the total number of syllables divided by total time, including pause, in seconds
Articulation rate (syllables/minute)	60 sec./min. times the total number of syllables divided by total time, excluding pauses, in seconds
Phonation/time ratio (percentage ratio)	Percentage ratio of time speaking to time to take the whole speech sample
Mean length of runs (number of syllables)	Average number of syllables between pauses (period of silence ≥ 250 ms)
Silent pauses per minute (number of silent pauses/minute)	60 sec./min. times total number of filled pauses (periods of silence > 250 ms) divided by the total time speaking in seconds
Mean length of pauses (seconds)	Mean length of all pauses (periods of silence > 250 ms)
Filled pauses per minute (filled pauses/minute)	60 sec./min. times total number of filled pauses (pauses filled with *uhm*, *mm*, *er*, etc.) divided by the total time speaking in seconds
Dysfluencies per minute (dysfluencies/minute)	60 sec./min. times total number of dysfluencies (repetitions, restarts, repairs) divided by the total time speaking in seconds
Pace (stressed words/minute)	Number of stressed words per minute
Space (ratio of stressed words/total words)	Proportion of stressed words per minute to total number of words

While earlier fluency researchers tended to use the term generally without specific reference to Segalowitz's framework, researchers are increasingly becoming more specific about which areas of fluency they are investigating. As can be seen above, researchers have a multitude of choices to make when deciding which fluency variables to analyze. Furthermore, even studies in which the same variables are investigated can differ in that decisions must be made for the minimum length of pause, and whether to focus on words or syllables. This situation makes it difficult to generalize the findings from one fluency study to another. Segalowitz (2010) has urged researchers to be more consistent with the measures they use and has argued that further investigation should reveal “a reasonably small set of cognitive processes that can be reliably associated with an equally reasonable small set of utterance fluency phenomena” (p. 51).

Longitudinal Studies of Fluency Development

One of the first studies that investigated fluency development over an extended period of time was Segalowitz and Freed’s (2004) investigation into the effects of study abroad on cognitive fluency development. In this study, the researchers sought to distinguish the differences in both cognitive and utterance fluency among 40 students studying Spanish as a foreign language either through a study abroad program or taking a foreign-language course at their home university. Utterance fluency data were obtained from extracts of student speech taken from oral proficiency interviews. The authors looked at four measures: speech rate, mean length of run with no silent pauses longer than 400 ms, mean length of run with no filled pauses, and the longest run containing no silent or filled pauses. To measure cognitive fluency, the students took a lexical access test to measure the reaction time and accuracy of word recognition skills, and an attention control test to measure reaction time and efficiency of cognitive processing. L1 cognitive fluency was also measured and controlled for in the analysis. Both the utterance and cognitive fluency measures were administered at the beginning and end of a 13-week academic semester. Overall, the results showed that the study abroad group made significant improvements in most of the utterance fluency measures, while the at home group made none, despite the fact that the groups did not show significant differences in pre-test measures. In contrast however, over the semester, both groups made improvements in cognitive fluency as measured by their word recognition reaction time scores, as no effect was found for the context of study.

The authors also looked at the degree to which the pretest cognitive fluency measures related to gains in oral fluency. Both pre- and posttest scores on the lexical access task were found to correlate positively with utterance fluency measures. Furthermore, utterance fluency gains significantly correlated with initial scores on the lexical access task, suggesting that utterance fluency gains might depend upon cognitive readiness, that is, there might be a threshold level of cognitive fluency necessary to engage in some communicative interaction. The scores on the attention control task were also significantly related to utterance fluency; however, this was a negative relationship, suggesting that as students began to gain more power in monitoring their speech, it might slow down. The authors concluded that while the different contexts lead to different changes in utterance fluency, the degree to which the students used their environment to improve their language skills might have been related to their initial levels of cognitive fluency. Therefore, the study abroad students with lower levels of cognitive fluency might have avoided situations that required them to process large quantities of language, which resulted in their cognitive fluency development not differing from their peers who studied at home. These findings indicate that potential relationships could exist between cognitive and utterance fluency. However, due to the intrusive nature of the computer-based testing of cognitive fluency, it is difficult to replicate such research methods in classroom settings when learners are interacting with one another in groups. No researchers have conducted major studies in which they have used a similar methodology with respect to speaking fluency.

These researchers also investigated fluency development in students studying French as a foreign language (Freed, Segalowitz, & Dewey, 2004). In this study, three groups were compared: students studying for 3-4 hours a week in their home country, students in an immersion program studying 3-4 hours a day, and students who studied abroad, studying 2-5 hours a day. In addition to utterance fluency measures, the researchers also calculated how much time the students spent using the target language outside of the classroom. The results showed that the immersion students made greater gains in fluency measures than the study abroad group, and the home country group made no gains at all. The authors explained this unexpected finding with reference to the amount of time spent using the target language outside of the classroom, as the immersion students spent significantly more time doing so than the study abroad students. Such findings support the arguments of practice and skill-acquisition theory, e.g. DeKeyser (2007b), which place a high importance on spending time on task. It should, however, be noted that this study did not include any cognitive fluency variables; therefore, these results should be interpreted cautiously, as the measures and analyses used are quite different from those employed in the other studies reviewed in this section.

A second series of studies investigated the link between cognitive and utterance fluency by combining several utterance fluency measures to infer cognitive fluency. As mentioned above, it has been argued that gains in automaticity and proceduralization imply more than a simple increase in the speed of production (Segalowitz, 2003). For that reason, several researchers have argued that increases in utterance fluency measures such as the mean length of run accompanied by stable pause times and phonation/time ratios suggest that language processing capacity, that is, cognitive fluency, has improved (Towell, 2002; Towell et al., 1996; De Jong & Perfetti, 2011). This is because speakers are able to produce longer and more complex speech using the same amount of time. The results of two studies of advanced learners of French showed that after completing study abroad sojourns of six months or longer, learners became more fluent, with greater gains seen for students with initially low levels of speaking fluency (Towell, 2002; Towell et al., 1996). While the lower level students in these studies made more gains than those of higher levels, they did not achieve the same degree of cognitive fluency as the higher proficiency students. Such findings suggest that fluency development could follow a power law of learning, where larger gains are made earlier in the learning process, and later development requiring far more time for improvement (DeKeyser, 2007).

In a more recent study, Mora and Valls-Ferrer (2012) compared the development of CAF measures - including a mean length of run and mean length of pause - in two contexts, study abroad and EFL formal instruction. This study is unique in that it investigated the effects of the two contexts on one group of 30 students over a two-year period, comparing the gains made in their initial study in an EFL context with the subsequent gains made after a three-month study abroad period. The results indicated that during the EFL study, the learners made no improvements in the fluency measures. However, after completing the study abroad program, significant gains were seen in most fluency measures. In particular, there were improvements in both the mean length of run and the mean length of pause, suggesting that the learners might have been able to proceduralize some of their lexico-grammatical knowledge whilst studying abroad.

The development of L2 fluency among immigrants enrolled in ESL classes has also been an issue of interest for researchers. A series of studies (Derwing, Munro, & Thomson, 2007; Derwing et al., 2009; Derwing & Munro, 2013) followed speakers from two cultural backgrounds, Chinese Mandarin speakers and Eastern Europeans who were native speakers of Slavic languages, that had immigrated to English-speaking communities in Canada. The initial study (Derwing et al., 2007), which was mainly concerned with the relationship between L1 and L2 fluency, found that while there were some significant correlations between L1 and L2 fluency in the initial stages of instruction, after the two-year period at which the study concluded, the connection between the L1 and L2 fluency measures had weakened and was non-significant for both groups. Both groups showed some significant improvements on their utterance fluency measures, but only the Slavic speakers were given significantly higher ratings in perceived fluency by expert raters. A follow-up study (Derwing & Munro, 2013) using data from the same group for a further five-year period showed that the Slavic speakers had continued to make gains in their rating scores, whereas the

Chinese speakers did not. The authors posited that individual difference factors such as willingness to communicate and confidence might account for this outcome.

There have also been quasi-experimental studies of cognitive fluency development in ESL contexts. De Jong and Perfetti (2011) investigated the 4/3/2 fluency training activity, during which speakers give three speeches under increasing time pressure. The researchers sought to determine the effects of repetition and examined 20 ESL students assigned to one of two conditions. The study was conducted over a period of four weeks. The experimental repetition group was instructed to speak about the same topic for each round of the 4/3/2 activity, while the control group was given a new topic for each round of the activity. The authors predicted that students in the repetition condition would make more gains in proceduralization, that is, an increase in cognitive fluency, than those in the control group. Furthermore, they predicted that these gains would also transfer to new topics, and to a greater degree for the experimental group. Following Towell et al. (1996), improvements in cognitive fluency were said to have taken place when there were increases in the mean length of fluent run while pause lengths and phonation/time ratios remained stable, as this indicates that speakers can produce more fluent language without an increase in planning time. The authors also distinguished between short- and long-term measures of fluency gains. Short-term fluency gains were theorized to occur when the speaker repeated information that had been produced in the recent past, that is, a previous round of the 4/3/2 activity, and is argued to be aided by the priming that results from producing language in previous utterances. In contrast, long-term fluency gains are argued to occur when speakers talk about a topic they have not discussed in the recent past. Gains in long-term fluency cannot depend on priming, and are therefore, the result of increases in proceduralization of productive language skills. The researchers analyzed changes in short-term fluency by looking at improvements across repetitions of speeches in the 4/3/2 activity, and long-term fluency by looking at improvements across novel two-minute monologue tests administered several weeks apart. The results showed that students in the experimental/repetition condition made improvements in long-term/cognitive fluency while those in the control group did not. The gains in cognitive fluency, however, followed two patterns. In the first, the mean length of fluent run increased while the length of pauses and phonation/time ratio remained stable, similar to the results found in Towell et al. (1996). In the second pattern, the mean length of fluent run remained stable, while the length of pauses decreased, and the phonation/time ratio increased. The authors argued that as speakers were producing the same amount of information with less planning time, this second pattern could also be argued to indicate an increase in cognitive fluency.

De Jong and Perfetti's (2011) study is important in that it investigates claims that the repetition of items is instrumental in developing both cognitive and utterance fluency. To follow up on their findings, they also analyzed the degree to which students repeated the same vocabulary across the three speeches in the 4/3/2 activity. For most of these activities, the number of repeated words correlated with a decrease in the phonation/time ratio, meaning speakers were filling more time with speech, in addition to a decrease in the mean length of pauses. This was the first major study to investigate the amount of language repeated in relation to fluency measures.

Other studies in ESL contexts have also shown development in cognitive fluency measures over a short period of time. Tavakoli, Campbell, and McCormack (2016) investigated the effects of fluency awareness-raising activities and strategy training on oral performance. This training included identifying fluency breakdowns in recorded monologues, practice of speaking strategies such as fillers and use of chunks, and out of class recording and analysis of speech samples. In their study, 37 students in an EAP program were split into two groups, the experimental group (n = 19) receiving the fluency treatment over a four-week period, whilst the control group (n = 18) studied general speaking and listening skills. The results indicated that the experimental group made small but significant gains in mean length of run with stable levels of pausing, whilst the control group did not, although they did make some gains in their spoken accuracy. The authors concluded that fluency development might be influenced by pedagogic interventions, such as those utilized in their study.

Studies of Fluency Development Activities

Prior to De Jong and Perfetti's (2011) study, improvements on student performance across speeches in the 4/3/2 task were investigated in two cross-sectional studies. While these studies were not as rigorously conducted as De Jong and Perfetti's research, they offer insights into how other aspects of the CAF framework might be influenced by fluency development activities.

Nation (1990) looked at whether the 4/3/2 technique improved language abilities, having six adult learners retell a personal experience, and one of those learners also repeated the activity with narrative and picture description tasks. There were two fluency measures: the number of words spoken per minute, and the number of hesitations, repetitions, and false starts per hundred words for each delivery. Accuracy and complexity measures were also included in this study. Accuracy was determined by the number of errors per 100 words. Fluency and accuracy figures were calculated by dividing the difference between the first and third speech by the figure for the first speech, then multiplying that figure by 100 to arrive at a percentage. Complexity was addressed qualitatively by examining how speakers reduced their content in subsequent speeches. The results showed that speech rate significantly increased while hesitations, false starts, and repetitions significantly decreased. The findings for accuracy were mixed, as four of the eight speeches showed improvements in accuracy whilst one remained stable and three subsequent speeches had more errors than the initial four-minute speech. Qualitative analysis showed that one speaker used more complex constructions in subsequent speeches whilst the remaining students omitted irrelevant details.

Arevart and Nation (1991) conducted a similar study with a larger sample size of 20 intermediate ESL students, mixed with partners of differing L1s. Similar to the above study, fluency was measured by words per minute and the number of hesitations per 100 words. Accuracy and complexity were not considered. For words per minute, results showed that changes ranged from an improvement of 62.1% to -1.4%, with significant average improvement rate of 21.5%, 18 words per minute. For the number of hesitations per 100 words, the results showed that changes ranged from -76.4% to 81.8%, with a significant average decrease in hesitations of 22% or 3.88 hesitations per 100 words.

While these studies also show improvement in utterance fluency, De Jong and Perfetti (2011) argued that such results can only be seen as indicators of short-term fluency development because improvements in fluency between speeches might be the result of syntactic or structural priming. Syntactic priming is the "phenomenon whereby the act of processing an utterance with a particular form facilitates processing a subsequent utterance with the same or a related form" (Pickering & Branigan, 1999, p. 136). Priming experiments have shown that when L1 speakers are asked to describe a situation, they can do so faster than when they have previously produced a sentence with a similar syntactic structure (Smith & Wheeldon, 2001). Because of the ability to speak faster when using the same or similar syntactic forms, it has been argued that syntactic or structural priming can facilitate speaking fluency because it reduces the cognitive load associated with producing an utterance (Pickering & Ferreira, 2008). In the two cross-sectional studies mentioned above, the lexical items and syntactic structures were primed in the first speech, making them easier for students to access in later speeches. For this reason, De Jong and Perfetti recommended longitudinal studies in which repetition activities are supplemented with additional activities where speakers discuss different topics, such as the monologue tests they used in their study, so that long-term fluency gains and proceduralization can also be determined.

Finally, in a recent series of studies (Boers, 2014; Thai & Boers, 2016), Boers and his colleague have questioned some of the claims made by researchers investigating fluency development, in particular, De Jong and Perfetti's (2011) claim that lexical overlap can aid language development. In the larger of the studies, Thai and Boers compared one group that completed an increasing time pressured 3/2/1 repeated speaking activity with a control group that completed an unpressured 2/2/2 repeated speaking activity. Consistent with previous findings, the results showed that learners in the increasing time pressure condition were able to make significant gains in their speaking fluency; however, in contrast, the students in the unpressured time condition made small, but significant gains in the accuracy of their language. Based on these findings, the authors claimed that encouraging learners to repeat monologues under time pressure could cause them to proceduralize incorrect language forms, as the lack of improvement in accuracy could be caused by learners repeating strings of incorrect language items in subsequent deliveries of their speeches.

There are however, several points that question whether this argument has been made prematurely. First, as the study was not longitudinal, there is no empirical evidence to suggest that task repetition activities that make use of time pressure cause students to automatize inaccurate language forms. Second, the authors report correlations among their variables, which show that there is no relationship between repetition and accuracy ($r = -.01$, $p > .05$), meaning that development in short-term accuracy was not related to the degree to which students used repetition of word strings. Finally, verbatim duplication, the measure which the authors used to measure lexical overlap only looked at language used in the first and final speeches. As De Jong and Perfetti's (2011) research shows, there is a much stronger relationship between the words used in all three speeches as opposed to two. Therefore, while the authors are correct in questioning whether fluency activities might also improve accuracy and complexity, as some researchers have suggested (Arevart & Nation, 1990; De Jong & Perfetti, 2011), the findings of the study do not support their claim that fluency development harms the development of complexity and accuracy.

Studies of Task Design and Fluency Development

Several researchers have looked at how levels of utterance fluency might be affected by task design using the CAF framework. While these studies are often cross-sectional or not specifically concerned with fluency development, they offer insight into which features of speaking tasks can have positive or negative effects on levels of fluency. One issue that makes it difficult to generalize findings from fluency studies is the difference among tasks used to elicit speech samples (Segalowitz, 2010). Studies that investigate task design and speaking fluency can make it possible to learn which speaking tasks are the best for helping students to improve their fluency levels.

One of the most cited studies is Bygate's (2001) investigation of task type and repetition within the CAF framework. The study involved 48 ESL students completing speaking tasks over ten weeks. There were two types of tasks; a narrative story retelling of a short video and an interview task involving picture descriptions. The students were split into three groups. All groups completed both tasks in a pretest on the first week of the study. The first experimental group then repeated the narrative tasks three times in weeks three, five, and seven. The second experimental group repeated interview tasks on the same weeks. The control group did not participate in any tasks during this time. At the end of ten weeks, all groups were then post-tested on the same two tasks. The materials for the pre- and posttests were exactly the same, while those for the practice tests in the interim sessions were different. Fluency was measured by the number of unfilled pauses per *t*-unit, accuracy was measured by the number of errors per *t*-unit, and complexity was measured by the number of words per *t*-unit. The first hypothesis was that narrative task would result more complex speech that was less fluent and accurate because of the cognitive load involved in retelling the story from the video. Second, it was hypothesized that task repetition would result in increases in fluency, accuracy, and complexity. Finally, it was predicted that the experimental groups would outperform the other groups on the task that they had practiced over the course of the study.

The results for the first hypothesis were the opposite of what was predicted. The interview tasks resulted in less fluent but more complex language forms than the predicted narrative tasks. The author argued that this could have been a result of having to respond to an interlocutor in real time, suggesting that monologue tasks can result in more fluent performances. For the second hypothesis, the results were complex. Repetition was related to significant improvements in complexity on both tasks, and fluency in the interview task. Changes in accuracy were not significantly related to repetition. The third hypothesis was not supported in that both experimental groups did not outperform the groups that did not practice their tasks between the pre- and posttests. These results showed that repetition can improve performance even when separated by a relatively large period of time (ten weeks). The results also showed that student performance can significantly differ according to the type of task they are performing. However, both Bygate and others (De Jong & Perfetti, 2011) have commented that one reason for the lack of differences between groups could have been the small number of measures used for each of the CAF variables. For this reason, they have stressed the importance of using a variety of measures until the most reliable ones have been determined.

Other researchers have analyzed the effects that task repetition could have on learners' attention when they speak. Fukuta (2015), studied the effects of task repetition by comparing the CAF measures of a group of Japanese university students (n = 14) who repeated the same picture description task, with those of another equal sized group who described different pictures in the two sessions. The author found that the accuracy and lexical variety measures in the experimental group were significantly larger than those in the control group. The findings are interpreted with Skehan's (1998) limited capacity model, as the author argued that by not needing to focus on conceptual issues, the speakers in the experimental condition were able to attend to lexico-grammatical matters.

Studies of Learner Reactions to Task Characteristics

While tasks can be designed to develop specific aspects of the CAF framework, learners do not always behave in a predictable manner. Murphy (2003) studied the performance of eight ESL students on three different tasks, each designed to develop a different aspect of the CAF framework in accordance with research findings. The purpose of the study was to investigate whether learners' performance in a specific CAF measure could be influenced by a task designed to promote that specific area of CAF, in accordance with criteria suggested by Skehan (1998). In this study, complexity was measured by the number of clauses divided by the number of AS units, accuracy was measured by the ratio of error-free clauses to the total number of clauses, and fluency was measured by the total number of words per second. The findings of the study showed that learners' performance generally deviated from what the research would predict. Only two of the eight students achieved their highest measure of accuracy on the task designed to promote accuracy. Fluency scores were relatively equivalent on the task designed to promote fluency in addition to the task designed to promote complexity. Student performance in complexity was the lowest on the task that was designed to promote it. While the author mentioned that the small sample size might have influenced the lack of consistency with previous findings, she suggested that while tasks can be designed in accordance with theory and research findings, individual learners might perform in unexpected ways. Such conclusions stress the importance of investigating all of the three CAF measures when investigating activities designed to promote a single aspect of them, such as the 4/3/2 speaking task that aims to develop speaking fluency.

While skill acquisition theorists emphasize that large amounts of practice are necessary for fluency development, few researchers have examined the quantity of student participation during EFL classes in relation to the development of CAF measures. Delany (2012) examined the participation of 37 Japanese EFL university students in two oral communication classes. An oral narrative task was used as a pretest and posttest at the beginning and end of the study, which took place over one semester. To measure student participation, the researcher conducted three discussions in which the entire class participated. Performance on the tests and class discussion was measured using CAF measures, in addition to the number of words spoken during the whole class discussion. To measure accuracy, the ratio of error-free clauses to the total number of clauses was calculated. Complexity was measured by calculating the ratio of clauses per AS unit. Fluency was determined by three measures: speech rate in terms of the total number of syllables produced, the mean length of run with .5 seconds as a cut-off point, and the mean number of hesitation phenomena per AS unit. The results showed that there were no correlations between any of the pre- and posttest gains in the CAF measures and the quantity of words spoken in class, suggesting that participation is not a significant predictor of development in CAF. These findings, however, should be taken with caution, due to the nature of the participation measure and the nature of the task itself. While the researcher argued that participation in the discussion was similar to that in regular classroom activities, participation in discussions of over 15 members are uncommon occurrences even for most L1 speakers, and students might have performed differently in smaller groups without the pressure of speaking in front of many peers. Furthermore, as the pre- and posttest contained the same question, task-repetition effects cannot be disentangled from language gains.

Pedagogical Frameworks for Developing Fluency

Although yet to be empirically tested, several researchers have offered pedagogical frameworks for developing fluency in instructed language settings. Several researchers have argued that fluency can be developed more quickly and effectively if learners have the opportunity to participate in activities that have been designed in accordance with theories of fluency development (Gatbonton & Segalowitz, 1998, 2005; Rossiter et al., 2010; Wood, 2001); however, there are few published materials developed with this purpose in mind. In a survey of both commercial learner textbooks and teacher resources, Rossiter et al. found that most texts did not include activities to develop learner fluency and argued that more attention needs to be directed toward this area. This is because the authors of several studies, such as Derwing et al. (2007) and Derwing and Munro (2013), have argued that a lack of fluency development activities might have been one reason for the lack of development in perceived fluency over extended periods of time. Frameworks for L2 teaching pedagogy such as Nation's (2001) four strands, allot a quarter of all instruction time to fluency activities where learners should focus on using the language they have already learned. Nation and Newton (2008) argued that without such provision, learners will not be able to use the language they are learning, rendering the effectiveness of instruction highly questionable. In recent years, there has been a development of pedagogical frameworks that have incorporated findings from cognitive science research into their principles with the purpose of improving speaking fluency.

One of the first attempts to create a framework for principled fluency development activities is Gatbonton and Segalowitz's (1988) work on creative automaticity. This term was created to refer to developing automaticity in respect to specific formulaic utterances associated with performing language functions rather than language structures. The authors specified five principles that materials designers need to consider when creating such activities:

- Activities should be genuinely communicative
- Activities should incorporate authentic aspects of real communication, e.g. require

listeners to experience the pressure of having to respond to questions in real time

- Activities should be focused on language functions used in the real world
- Activities should require learners to use short formulaic utterances when communicating
- Activities should be inherently repetitive, in that learners can repeat the activity meaningfully with different partners

When creating activities that incorporate these principles, Gatbonton and Segalowitz outlined a two-step procedure. The first step involves creating an activity that creates an inherent need for learners to use formulaic utterances, which they specify as the objective of the lesson. The second phase consists of more controlled practices that aim to have students use the utterances in meaningful communication and interactions with their classmates.

This framework was later expanded into the ACCESS fluency development methodology (Gatbonton & Segalowitz, 2005). Incorporating the above principles, the authors created a three-phase process for developing spoken fluency. The first two phases are similar to the two phases outlined in Gatbonton and Segalowitz's (1988) work on creative automaticity. The first phase, the creative automatization phase, is designed to elicit the need to use specific utterances in meaningful tasks such as role-plays and problem-solving. The second phase, the language consolidation phase, aims to build learners' abilities to use target utterances through the use of fluency, accuracy, and grammatical development tasks. The third and final phase is the free communication phase which is designed so that learners can practice the target utterances in different contexts by participating in activities similar, but different, to those employed during the first phase. The authors argued that by including the ACCESS model as a component in a language-teaching program, learners can build automaticity and fluency with the target language.

Segalowitz (2010) listed four criteria that he argued need to be taken into account when designing fluency development activities. These criteria are taken from a cognitive science perspective, with the goal of enhancing “the cognitive fluency of the processes underlying L2 speech production” (p. 173). The four criteria are activities that allow for transfer appropriate processing, activities that are genuinely communicative, activities that provide affordances for pursuing linguistic goals, and activities that are inherently repetitious. By taking these four criteria into consideration when designing fluency activities, Segalowitz argued that learners will be able to build their cognitive fluency and automaticity when producing speech.

The first of Segalowitz’s criteria is that the activity must involve transfer appropriate processing (TAP). This criterion comes from the TAP principle that was developed from a series of studies involving how fluently individuals can retrieve information from memory. Segalowitz (2010) cited results from studies that have suggested that the degree of fluency with which individuals can recall information depends on the degree of similarity to the situation with the context of the time of learning the information concerned. This issue offers many challenges for language teachers, as classroom activities need to be as contextually similar to real world activities. If this condition is not met, learners will not be able to use the skills they have learned when communicating with others. With regards to this challenge, Lightbown (2007) argued that learner needs must be determined so that the distance between the kinds of processing used in the classroom with those outside of the classroom is minimized.

The second criterion, genuinely communicative activities, comes from L2 teaching pedagogy and L1 usage-based theories of language acquisition. CLT advocates have long argued that classroom activities need to be primarily focused on meaning rather than the study of language forms (Ellis, 2008; Nation, 2001), under the assumption that learners best acquire an L2 through engaging in communication using it. Segalowitz noted that while many issues remain with this principle, the need for genuinely communicative activities is supported by recent usage-based language acquisition theories, such as Tomasello (2003). Tomasello stated that when communicating, interlocutors engage in joint attention processes through monitoring each other's attention and paying attention to the objects or events being discussed. Furthermore, they also engage in intention reading by attempting to determine each other's social intentions. Therefore, learners need to engage in these two processes if they are to acquire the skills necessary for successful L2 communication. The decontextualized and overly form-focused activities seen in many language classrooms and textbooks do not require learners to take a fully communicative situation into account, as they do not fulfill communicative goals in the same sense that real-world communication does. If teachers want to successfully develop their learners' oral fluency, Segalowitz has argued that they must provide activities that require them to engage in the processes of joint attention and intention reading, so that they are genuinely communicative.

Segalowitz’s third criterion, linguistic affordances, concerns the relationship between the learner and the environment with which they are using the target language and is taken from an ecological perspective on SLA (Van Lier, 2000). Van Lier incorporated the term *affordances* into SLA from the field of psychology. Gibson (1986) created this term in his theory of visual perception. In its original sense, it refers to how organisms create action through their perception of what the features of the environment offer them. For example, a spider might see a leaf in a forest as a means of shelter from the rain, whereas a caterpillar might see the same leaf as food (Van Lier, 2000). Transferring the term to a linguistic environment, being a fluent speaker requires an individual to be aware of a language’s possibilities and limitations, and taking those into account, being able to use the language to achieve their communicative goals (Segalowitz, 2010). Therefore, in addition to knowledge of language forms, speakers also need to be aware of how to manipulate language to produce speech acts, such as apologizing or asking questions. If learners are unable to retrieve or produce the speech acts necessary for achieving their communicative goals, they will not be able to communicate using the target language. Segalowitz argued that fluency development activities need to provide opportunities for linguistic affordances so that learners can develop their skills in achieving communicative goals. This skill is important because if utterances are not spoken fluently, learners might not be able to perform the speech act that they wish to carry out.

The final criterion is that fluency development activities need to be inherently repetitive, as without opportunities to repeatedly practice language that has been previously learned or used, automaticity cannot develop. Segalowitz (2010) pointed out that repetition in this sense must be meaningful, and not simply repetition for its own sake, such as form-focused drills common in Audiolingual methods of language teaching (Richards & Rodgers, 2001). Rather, repetition must be a means of achieving some kind of communicative goal. Maurice's (1983) 4/3/2 activity meets this criterion in that learners must relate the same information to different speakers, giving them two further opportunities to repeat previously produced language in a genuinely communicative context.

There have also been other frameworks offered to suggest means of fluency development. Although these frameworks are less detailed than those of Gatbonton and Segalowitz, they share many features while also introducing additional principles or areas to consider. Wood (2001) offered a four phase fluency course based on findings from theories of automaticity and formulaic language. The first phase is unique in that it utilizes listening activities to draw students' attention to the features of fluency in native speaker discourse. In addition to formulaic utterances, learners' attention is also directed toward features of speech such as pausing and hesitations, to raise learners' consciousness of how they are used by native speakers of the L2. The second phase is the automatization stage where learners shadow the recording in a language laboratory and complete communicative activities that require them to use the formulaic language being studied. Following this is the practice and production stage where learners complete a 4/3/2 activity. Finally, learners complete a free talk stage where they engage in an extended discussion on a topic, followed by a discussion in which learners evaluate the fluency with which they could communicate.

Nation and Newton (2008) suggested five principles for creating fluency development activities. First, the activities must be easy to ensure that all of the language required for completing the task is within the learner's level. Second, the activities must be meaning-focused so that they have a clear outcome. In relation to this point, the authors argued that while the 4/3/2 activity might not be goal-oriented in the sense that more authentic tasks might be, it is still meaning-focused because learners are required to share their opinions or experiences with an audience. The third principle is time pressure, which is argued to be an important means of encouraging learners to perform at a higher level than that of their usual language use. For example, time limits, such as the reduction of one minute from each round of the 4/3/2 activity, can prompt learners to speak at a faster rate than they usually do in classroom activities. The fourth principle, planning and preparation, is suggested as a different means of encouraging higher levels of performance. Research has shown that learners can make gains in fluency and complexity when after completing rehearsal planning activities, and while gains in accuracy are less common, this can be addressed by additional interventions (Ellis, 2005, 2009). The final principle is that of repetition, which is argued to build fluency in items and sequences used. Nation (1989) described a variety of activities that can be adapted to fit the above criteria.

While the above frameworks and Segalowitz's four criteria are closely connected to research carried out in the field of cognitive science, to date, the only aspect of fluency development pedagogy that has been empirically tested is that of inherent repetition, in De Jong and Perfetti's study of the 4/3/2 activity. While the results of this study were positive, researchers have yet to show how the adoption of these other ideas and principles can assist in the development of speaking fluency. Commenting on Gatbonton and Segalowitz's creative automatization principles, Bamford (1989) argued that it is challenging for teachers to develop materials that incorporate many different principles, and the authors conceded that the success of such methodologies could depend on the skill of the teacher (Gatbonton & Segalowitz, 1989). Furthermore, some of the principles might be difficult to apply in some educational settings. To account for transfer appropriate processing in language classes, course designers must be aware of student needs. This can be particularly challenging in many EFL settings where learners are taking compulsory English classes, as the students themselves might not yet have decided upon how, or if, they will use English in their future.

Fluency and Individual Differences

While the majority of L2 speaking fluency research has been focused on the theorized cognitive processes underlying productive speaking skills, several prominent researchers have argued that there are social and individual difference aspects to L2 speaking fluency as well (Segalowitz, 2010; Wood, 2010). Segalowitz, Gatbonton, and Trofimovich (2009) argued that individual difference variables such as motivation can affect the amount of experience that a speaker has using the L2, which can affect "a number of *pscholinguistically* relevant cognitive variables … that normally would contribute to L2 development (p. 174, italics original). After examining the results of several small studies, the authors concluded that motivation affects the ways learners select communicative experiences, and could therefore, affect their level of L2 attainment. Wood (2010) singled out four individual difference variables he argued can affect L2 speaking performance: language anxiety, self-efficacy, voice factors concerning issues of power-relations and speaking, and social identity. However, there has been little published research with regard to how these factors affect L2 speaking fluency or its development. Two further individual difference variables that could affect L2 speaking fluency are extraversion and willingness to communicate (WTC) because of the central role placed on output in many theories of SLA. Therefore, given that these two individual difference variables might influence the degree to which a learner uses an L2 to communicate, learners with higher levels of extraversion or WTC might be able to develop speaking fluency at a faster rate.

Extraversion. Extraversion is a hypothesized dimension of personality, often measured in relation to introversion, which is a component of the Big Five personality trait model created by Goldberg and comprehensively operationalized by Costa and McCrae (Dörnyei & Ryan, 2015). This dimension of personality is related to whether a person tends to focus on the outside world around them, or the inner world of their thoughts, and which of these worlds is a source of their energy (Dörnyei & Ryan, 2015). In a seminal study, 31% of good language learners stated that they believed extraversion to be helpful in develop oral speaking proficiency (Naiman, Frolich, & Stern, 1996). SLA researchers have attempted to understand the relationship between language proficiency and extraversion, although the results have been mixed and difficult to generalize due to a variety of methodological issues and the different instrumentation measures used across the studies (Dewaele & Furnham, 1999).

However, despite some contradictory findings, there is some degree of consensus that extraversion can benefit the social types of learning processes that occur in communicative language classes (Dörnyei & Ryan, 2015), although it has been noted that the majority of these studies have used paper-based test scores as measures of language proficiency, rather than the actual output produced by learner, questioning the validity of the findings (Dewaele & Furnham, 1999).

There are however, two published studies looking at the relationship among extraversion and CAF measures of spoken output, however, with regards to speaking fluency, the results are contradictory. Dewaele and Furnham (2000), obtained several CAF measures in both formal and informal situations from 20 Flemish university students learning French. They found several significant correlations among all aspects of CAF and levels of extraversion, especially in formal settings, where the more extraverted students significantly outperformed the more introverted students on four of their six measures. They argued that the effects of extraversion might be more pronounced when learners are given relatively demanding tasks in formal settings. In such settings, more introverted speakers could be overwhelmed by having to perform a speaking task, and this could result in less and/or lower quality output. However, a follow-up study of 25 Flemish learners of French and English revealed no significant effects between fluency and extraversion, which led the authors to suggest that extraversion may only influence speaking fluency in highly formal settings, as introverted students might not be inhibited by the relatively relaxed settings of a classroom (Van Daele, Housen, Pierrad, & Debruyn, 2006). To date there appears to be have been no further studies investigating the relationship between CAF variables and extraversion. Therefore, there is still a lot to be learned about the relationship between these personality variables, and whether or not more extraverted students can develop their speaking fluency in a second language at a faster rate than their more introverted peers.

Willingness to communicate. Willingness to communicate (WTC) is concerned with the degree to which L2 learners are willing to initiate communication with other interlocutors. While people may have relatively stable levels of WTC when speaking their first language, researchers have argued that the case is more complex for second language learners (MacIntyre, Baker, Clement, & Conrod 2001). In a summary of recent research on WTC, Dörnyei and Ryan (2015) note that the majority of published WTC research has investigated the specific effects that a learning context might have on WTC. These studies, have looked in depth at how WTC can be influenced by factors and or changes in the classroom (see Kang, 2005; Cao & Philp, 2006), and while they have revealed some insights into how individual levels of WTC are often dynamic in nature, there is still much to be learned about how levels of WTC might affect the development of speaking skills at a more general level.

An individual's level of WTC might have implications for the rate at which they develop their speaking fluency because learners who more frequently use the target language will have more speaking experiences. As a result, they might be able to increase the speed at which they automatize features of the L2. However, to date, the majority of WTC researchers have investigated students' reported frequency of L2 use in both real and/or imagined situations rather than measuring their actual use, therefore there is a lack of empirical evidence to show any relationship between speaking development and WTC. One study however, did find a significant correlation between high school students' levels of WTC with their amount of participation in communicative activities (Dörnyei & Kormos, 2000), suggesting that this is the case, although the very small sample size used in the study makes generalization of the results difficult. Furthermore, a recent study found that learner's self-reported levels of WTC predicted their use of language in the classroom (Munezane, 2014). This study had over 300 participants, and found a direct link between WTC and the number of words used in a class discussion. To date, there have been no major studies of WTC and the development of language features in the CAF framework.

The relationship between extraversion and WTC. A recent study looked at the relationship between personality factors, including extraversion and WTC (Oz, 2014). In a study of 168 Turkish EFL learners, Oz found a significant relationship of a moderate size between the two variables. Furthermore, extraversion had the strongest relationship to WTC of all the five personality traits. This leads the author to conclude that students who have higher levels of extraversion are more inclined to use the L2 to communicate. These two individual difference measures, therefore, could predict student performance and/or levels of engagement in language classes. In foreign language settings where opportunities for language use outside of the classroom are minimal, students with higher levels of extraversion and WTC might develop oral fluency faster than their peers with lower levels of these measures because of additional time they spend speaking and using the L2. It still remains unseen, however, whether there is a relationship between these individual difference variables and speaking development.

First language fluency. Finally, several studies have shown that first and second language fluency levels have a significant relationship, prompting Segalowitz (2010) to caution that by not accounting for levels of L1 fluency, L2 fluency researchers might introduce unwanted noise into their data. In their study, Derwing et al. (2009) found stronger relationships between L1 and L2 levels of fluency at lower levels of proficiency. They did not however, find any relationship between L1 fluency and L2 fluency development, leading them to question whether fluency is a language-independent trait. De Jong, Groenhout, Schoonen, and Hulstijn (2013), however, concluded that to some extent, L2 fluency can be more accurately measured when L1 fluency is taken into account. The results of their study of L1 English and Turkish speakers of Dutch showed that L1 measures could account for between .21 to .57 of the variance in fluency measures. However, for most of the measures, they found that both the L1 corrected and uncorrected fluency measures were equal predictors of L2 vocabulary knowledge. It should be emphasized the populations in these studies were quite different; Derwing et al.'s study focused on Slavic and Chinese immigrants living in Canada, while De Jong et al.'s study focused on English and Turkish immigrants living in Holland. Both of these studies conclude that more research in this area is needed to determine the extent to which levels of L1 fluency could influence L2 oral performances.

Summary of Research of Second Language Fluency

The term *fluency* can be interpreted in many different senses, and there are even several differing technical definitions of the word. While all definitions generally refer to temporal dimensions of speech production, Segalowitz has argued that the meaning of fluency differs depending on how it is being measured, resulting in three kinds of fluency: cognitive, utterance, and perceived. Cognitive fluency, which describes the speed with which language is processed in the mind, is most often measured by time reaction software. Utterance fluency, which describes the quality of speech produced in a performance, can be measured in a multitude of ways, the most common being objectives measures of the quantity of speech produced and the length of the pauses taken during the performance. Finally, perceived fluency refers to expert ratings of the quality of speech produced by a speaker, and is measured by various rating scales.

Due to the difficulty of measuring cognitive fluency in second language settings, several researchers have argued that cognitive fluency development can be inferred by looking at several utterance fluency measures together (De Jong & Perfetti, 2011; Towell, 2002; Towell et al., 1996). These researchers have looked at the relationship between the mean length of run and mean length of pause, and have argued that if one of these measures increases over time while the other remains stable, cognitive fluency has improved. Several studies have shown that study abroad programs (Towell, 2002) and fluency development activities (De Jong & Perfetti, 2011) can lead to improvements in cognitive fluency. In addition, there have been several cross-sectional studies of fluency development activities, the results of which suggest that tasks involving repetition can have strong effects on utterance fluency (Arevart & Nation, 1991; Nation, 1990). It is important to note, however, that such results might be the result of priming processes, hence, De Jong and Perfetti (2011) labeled such developments as short-term fluency.

Insights into L2 speaking fluency development can also be learned from looking at the CAF literature, which takes a more general view of SLA and investigates complexity and accuracy variables in addition to fluency measures. Some of these studies which have had students repeat the same task after a certain period of time has elapsed have shown effects on language development for task repetition (Bygate, 2001). However, the findings of such research must be interpreted cautiously, as results from studies such as Delany (2012) and Murphy (2003) have shown that students' speech performances can be unpredictable.

Finally, while yet to be comprehensively investigated, several researchers have created pedagogical frameworks focused on developing speaking fluency. These frameworks have been based on theories of speech production and automaticity. In one of the most comprehensive, Segalowitz (2010) based his approach around four criteria to be applied to fluency development activities, namely they must: allow for transfer appropriate processing, be genuinely communicative, provide affordances for pursuing linguistic goals, and be inherently repetitious. Furthermore, Segalowitz has argued that a successful theory of L2 fluency development also must take individual differences and other social factors associated with language learning into account, although to date, there appears to be few investigations of such issues.

Gaps in the Literature

Although the number and sophistication of research into L2 speaking fluency has increased in recent years, a review of the literature shows there is still much to be learned about how speaking fluency develops for L2 speakers. This is especially the case for EFL learners who have limited contact with speakers of English. This study is designed to address four specific gaps in the literature; the development of speaking fluency in EFL classes, the relationship between student performance in classroom activities and fluency gains, the relationship between individual differences and fluency gains, and an investigation of how EFL students' speaking fluency develops in respect to the accuracy and complexity in their spoken language.

The majority of fluency development studies have taken place in contexts were the learners were studying in a country where they had access to speakers of the target language. There have been few detailed studies that have investigated the degree to which EFL students can develop their speaking fluency in foreign language classrooms. Furthermore, in ESL settings, there has been only one longitudinal study of fluency development activities and gains in long-term fluency and proceduralization (De Jong & Perfetti, 2011). However, the students in this study were of a high-intermediate proficiency and were enrolled in an American university. Whether fluency development activities can benefit EFL learners of lower levels of proficiency remains to be determined.

Second, several prominent fluency researchers (Gatbonton & Segalowitz 1988, 2005; Wood, 2001) have argued that learners need to be presented with formulaic language and given opportunities to use them in relatively controlled practice activities that encourage development of creative automaticity. It is unclear, however, whether students who use these phrases more than their peers make larger gains in fluency, as skill acquisition theory would suggest. Wood (2010) concluded that use of formulaic language facilitated fluency development; however, this study was of performance in narrative retellings and was not focused on a set of formulaic phrases that had been specifically chosen by course designers. Whether learners' fluency development can be enhanced by their use of formulaic phrases has yet to be investigated.

In relation to this point, research on fluency development has not investigated whether quantity of speech in addition to repetition of vocabulary can affect rates of growth. To date, only the study by De Jong and Perfetti (2011) has addressed the issue of repetition; however, the researchers looked at one fluency activity, whereas students might be repeating more of the same language in activities conducted during earlier stages of the lesson. Furthermore, while theories of skill acquisition state that learners need to engage in a large amount of practice to develop their speaking fluency, few researchers have tracked the quantity of student speech in a variety of different activities. While Delany (2012) investigated oral participation and proficiency gains, his study analyzed performance in whole class participation in which the teacher was a participant. It is unclear whether students who participate more in small group work without any teacher intervention make larger gains than their more reticent peers, as skill acquisition theory suggests.

Third, little is known about how personality measures of individual difference constructs are related to fluency development. While several researchers have suggested that extraversion and willingness to communicate could be related to more proficient L2 speakers, to date, there has been no empirical investigation of how these factors influence the development of speaking fluency.

Finally, fluency development activities have not been studied in relation to longitudinal gains in accuracy and complexity. Several researchers have shown that certain learners improve their accuracy and complexity as they complete the 4/3/2 speaking activity (Arevart & Nation, 1991; Nation, 1989). However these studies have been cross-sectional with very small sample sizes. It is unclear what relationship that potential gains from fluency development activities could have with accuracy and complexity. As it has been argued that fluency should not be studied in isolation to accuracy and complexity (Ellis & Barkhuizen, 2005), the effect of fluency development activities on other areas of spoken language require further investigation.

Purposes of the Study

In this study I seek to investigate the degree of variability in EFL students' fluency development over several points in one academic semester. To date, most published researchers of fluency development have collected data at long-term intervals ranging from nine months to one year (Derwing & Munro, 2013; Derwing et al.; 2007; Lennon, 1990; Mora & Valls-Ferrer, 2012; Towell, 2002; Towell et al., 1996). While several studies that have investigated fluency development over an academic semester, one of these studies did not include temporal data (Du, 2013), and other studies had only pre- and posttest measures (Freed et al., 2004; Segalowitz & Freed, 2004). One study included three measurements over four weeks (De Jong & Perfetti, 2011); however, these measurements were taken from high intermediate level ESL students studying in an English-speaking university. How speaking fluency develops for lower proficiency EFL students using a higher density of observations has not been investigated.

The second purpose of this study is to determine which factors are related to development in both short- and long-term fluency. While De Jong and Perfetti (2011) have shown that repetition is associated with gains in fluency, there have been few investigations of how other factors that have been theorized to influence fluency development. In addition to investigating repetition and fluency development, this study also includes the quantity of language produced in class in addition to the use of directly taught formulaic phrases. Furthermore, I investigate whether individual difference factors, such as extraversion, willingness to communicate, and first language fluency, are associated with fluency gains.

The third purpose of this study is to investigate fluency gains in relation to accuracy and complexity in the CAF framework. To date, the only study that has looked at fluency development activities longitudinally has not addressed this issue, and cross-sectional studies have produced unclear findings. In this study I investigate whether students who make gains in fluency also build their speaking accuracy and complexity, in addition to investigating whether students who do not improve their speaking fluency do so in other areas of the CAF framework.

The final purpose of this study is to investigate the degree to which expert ratings of perceived fluency correlate with temporal measures of utterance fluency. Several researchers have investigated this question (Derwing et al., 2009; Lennon, 1990); however, it is still unclear which features of utterance fluency are most related to perceived fluency. In this study I attempt to determine which aspects of utterance fluency are most salient to expert raters.

Research Questions

The following research questions are investigated in this study.

1. To what degree do EFL students' levels of spoken fluency develop over one academic semester (change in long-term fluency)? To what degree does this development covary with in-class performance in communicative activities and individual difference variables?

2. To what degree do EFL students' levels of spoken fluency develop over a 3/2/1 speaking task (change in short-term fluency)? To what degree does this development covary with in-class performance in communicative activities and individual difference variables?

3. What relationships exist among development in complexity, accuracy, and fluency?

4. To what extent are expert ratings of perceived fluency related to temporal measures of utterance fluency?

Conclusion

In this chapter I reviewed the theories and relevant research related to the development of second language speaking fluency. Levelt's (1989) model of speaking was introduced which is largely agreed upon as the most comprehensive explanation of speech production (Kormos, 2006). The two major theories of automaticity, rule-based (Anderson, 1983) and instance-based (Logan, 1998) were then discussed as potential explanations of how speech can be produced and processed at a rapid rate, which Kormos (2006) has suggested, might account for different kinds of automatized knowledge. Formulaic language was then discussed as one tangible result of automatized language knowledge, which has been argued to be necessary for fluent speech in a second language (Wood, 2010).

With regards to second language fluency, a summary of the measures used to quantify fluency was given (Segalowitz, 2010), followed by a survey of the research in L2 fluency development. Most researchers seeking to determine the degree to which fluency has developed have done so by comparing changes in the mean length of run with those in the mean length of pause (Towell et al., 1996). While there have been very few longitudinal studies of fluency development in instructed settings a recent study suggests that repetition of the same lexical items may drive fluency development (De Jong & Perfetti, 2011).

I then discussed pedagogical approaches to developing L2 fluency in instructed settings, in addition to some of the factors that could affect its development, such as classroom performance factors, in addition to three individual difference variables: extraversion, willingness to communicate, and L1 fluency.

Finally, I identified several gaps in the literature, explained the purpose of this study, and summarized the research questions. In the next chapter, I describe in detail the methodology used to answer these questions.

CHAPTER 3

METHODS

In this chapter, I explain the methods used to answer the four research questions in this study. First, I describe the participants and the materials used in the classes in which this research was conducted. Next the procedures for collecting data are explained, followed by the design of the study. The methods for obtaining measures from the spoken data are then described, followed by the means of analysis. Finally, a chapter summary is provided.

Participants

The participants in this study were 32 first-year Japanese university students in four classes of between seven to nine students each. All students were studying at the same co-educational university, which is acknowledged to be a relatively prestigious university in central Tokyo, as the *hensachi* score for all of the university's eleven colleges were above 50. The *hensachi* is a score given to each department of a university based on how far from the mean an average student scored on the national Center Test and the college's entrance examination. A score of 50 represents an average student scoring at the mean, 60 indicates one standard deviation above the mean, and 70 indicates two standard deviations above the mean. There were 15 male and 17 female students. The students were selected for this study on the basis of being enrolled in my classes for the semester in which this study was conducted. In terms of general English proficiency, one class was of a high beginning level, one of low intermediate, one of intermediate and one of high intermediate. The high beginning level class were economics majors, a college with *hensachi* rankings ranged from 60 to 62.5, the low intermediate class were sociology majors, a college with *hensachi* rankings from 60 to 65, the intermediate level class were law majors, a college with *hensachi* rankings from 55 to 62.5, and the high intermediate class were non-English humanities majors, a college with *hensachi* rankings from 55 to 62.5. Prior to being admitted to the university, all students had completed a minimum of six years of formal education in junior and high school, including English classes. The participants were grouped into discussion classes based on their major field of study, and then ranked in terms of their scores on a 30-minute listening section of the TOEIC test (see Table 2). As there was no speaking component to the placement test, the classes were of relatively mixed abilities when engaging in speaking activities. All students had to take three compulsory English classes in each semester of their first year of university. The university's first-year English curriculum seeks to develop students' English abilities so that they can be active in international settings, and the majority of the courses are therefore focused on developing students' productive skills in English. In both semesters, the students took two courses that met weekly for one 90-minute lesson over two academic semesters; discussion and presentation. In alternating semesters, students also took one 90-minute writing class, or an e-learning course focused on vocabulary, listening, and grammar. Among the classes, the discussion class was unique in that there were only seven to nine students per class in order to maximize student talking time. The writing and presentation classes had an average of 20 students per class, while the e-learning class was held in computer laboratories with several large classes completing activities whilst being supervised by one instructor. The students in this study had all taken and passed the first semester discussion class.

Before beginning data collection, I completed an online version of the CITI (Collaborative Institutional Training Program). A consent form was designed to reflect the principals outlined in the CITI. All participants were informed of the purpose of the study and signed Japanese translations of the consent forms. Students were assured that their privacy would be protected, that the study would not affect their grades, and that they were free to withdraw from the study at any time. All students gave their consent to be included in the study. The deputy director of the program also gave permission for the study to be conducted.

Table 2. *Descriptive Statistics for the TOEIC Placement Test*

	Listening		Reading		Total		
Class	*M*	*SD*	*M*	*SD*	*M*	*SD*	*N*
1	382.14	26.75	296.43	53.83	678.57	67.19	7
2	278.13	2.59	240.01	74.60	518.13	74.69	8
3	236.25	3.54	259.36	50.32	495.63	47.92	8
4	117.78	97.37	115.56	86.53	236.67	38.16	9

The discussion class in which this research was conducted met once a week over two 14-week semesters in the students' first year of study. All discussion classes were taught under a set curriculum with the main goal of developing the students' spoken fluency as well as communicative skills. The course had an in-house textbook and teacher's manual that required the instructors to provide a minimum of 50-55 minutes of student-to-student interaction during each class, including one 10-minute and one 16-minute discussion during which students discussed a topic in groups of three to four without teacher intervention. While the textbook topics were identical across all levels, there were slight differences in the length of the homework readings, the number of formulaic phrases to be learned, and the complexity in the wording of the discussion questions in each level. The syllabus was organized around a two lesson topic cycle, and was based around social issues that first-year university students could discuss without reference to other sources. Topics included education, the environment, media, traditional vs. popular culture, and gender. This criterion was chosen to ensure that the students could focus on developing their speaking fluency by mostly using vocabulary items and language forms that they were already familiar with, one of Nation's (2001) conditions for a fluency development activity. The textbook questions and topics were revised with feedback from full-time instructors at the end of each semester to ensure that questions were relevant to student interests and capable of enabling students to generate extended discussions. At the time the study was conducted, the textbook was in its fourth edition, and most of the teaching staff felt that the topics were conducive for this purpose. The formulaic language taught in the course had been adapted from direct approaches to teaching conversational skills (Dörnyei & Thurell, 1992; Kehe & Kehe, 1994, 1998, 2003; Richards, 1990), with the criteria that learners would be able to use each phrase in every class discussion, as it was hoped that by constant use of the phrases, students' interactive skills would improve. In accordance with Segalowitz's (2010) principles of fluency activities, the formulaic phrases and the corresponding functions they performed were selected using the criteria of whether they were general enough for students to be able to use them in most communicative situations. This approach represented an attempt by the course designers to enable ease of repetition and to encourage students to make the most of the linguistic affordances offered in an interactive discussion in which peers were expected to contribute equally and share their views on an issue. Furthermore, the formulaic phrases were taught from both listener and speaker roles, to encourage students to be more active listeners during the discussions. While the majority of time spent in discussion class activities was devoted to meaning-focused communicative activities, the formulaic language phrases (See Appendix A for the full list of formulaic phrases) were directly taught and practiced under controlled and semi-controlled conditions.

The students were assessed on their attendance and participation in regular classes (70% of the final grade) and their performance on three group oral discussion tests (30% of the final grade). For regular lessons, students' grades were assigned holistically by the instructor, who took specific notes of student performance during activities. In test lessons, grades were assigned by quantifying the amount of student interaction in a 16-minute discussion with their peers. See Table 3 for an overview of the 14-week semester. During the first class of each semester, the assessment policy was explained to students, and they were also given a bilingual student handbook explaining the grading criteria in significant detail.

Regular lessons followed a uniform routine each week (see Table 4), beginning with a short quiz on the topic of the reading (see Appendix B for a sample quiz). Following this, the students engaged in form-focused practice of communicative functions, that is, the formulaic language from Appendix A. Each semester, six regular

Table 3. *Topics and Assessment for the Second Semester Discussion Class*

Lesson	Topic	Assessment
1	Introduction	Attendance only
2	Language 1	Attendance and Participation
3	Language 2	Attendance and Participation
4	Fashion 1	Attendance and Participation
5	Fashion 2	Discussion Test 1
6	Media 1	Attendance and Participation
7	Media 2	Attendance and Participation
8	Globalization 1	Attendance and Participation
9	Globalization 2	Discussion Test 2
10	Human Rights 1	Attendance and Participation
11	Human Rights 2	Attendance and Participation
12	Gender 1	Attendance and Participation
13	Gender 2	Discussion Test 3
14	Wrap-up	Attendance only

lessons (Lessons 2, 3, 6, 7, 10, and 11) focused on one of these communicative functions, while, three lessons functioned as review lessons prior to testing (Lessons 4, 8, and 12). During this time, the students completed short communicative drills that required them to use the lesson's target phrases. These drills were based around controlled cloze activities that students completed in pairs, followed by semi-controlled practice activities where students could ask and answer simple questions using the lesson's target language. A sample set of these activities can be seen in Appendix C). Following these drills, the students completed two extended discussions, both preceded by short preparation activities (see Appendices D and E for a list of these activities. The first discussion lasted an average of two and a half minutes per participant (usually 10 minutes for four students), and the second discussion lasted an average of four minutes per student (usually 16 minutes for four students). Following the second discussion, the students completed the 3/2/1 task and were finally given teacher-fronted feedback and short formative assessment activities at the conclusion of the lesson.

Table 4. *Regular Lesson Plan*

Time (min.)	Activities
1-5	Quiz
6-25	Function phrase presentation and practice (communicative drills)
26-40	Discussion 1 preparation, discussion 1
41-65	Discussion 2 preparation, discussion 2
66-85	3/2/1 Fluency
86-90	Feedback and formative assessment

The lesson plan for the discussion test was considerably different in that the first half of the lesson was focused on practicing previously taught functional phrases and activities to generate content for the test itself. The second half of the lesson consisted of the test, followed by teacher-fronted feedback. Because of the different nature of these lessons, student interactions in the three discussion test lessons were not included as part of this study.

Materials

Textbook-Based Activities

The textbook-based activities—the quiz, function practice, and Discussion 1—that took place before the second discussion were not included in this study because during that time, the students had access to their books that contained the functional phrases, in addition to the homework readings that contained information regarding each topic. During much of this time, the instructor directed students' attention toward the phrases, which were both written on the class whiteboard and in the textbook (see Appendices A and C for sample phrases and activities). Because speech production could have been a result of students directly reading out of the textbook rather retrieving the language to produce the speech themselves, the data from these activities were excluded from the analyses.

Discussion 2 Activities

As mentioned above, the Discussion 2 activities from the in-house textbook were specifically developed for the course. The Discussion 2 activities used in this study were taken from Lessons 3, 4, 6, 7, 8, 10, 11, 12, and 14. These discussion activities consisted of two stages: a preparation activity and the discussion itself. The preparation activities consisted of four short prompts that students had to agree or disagree with (see Appendix D for the full list of preparation activity prompts). There was a time limit of two minutes for each prompt, which was measured using an electronic timer. Each prompt was posted to the wall in the corner of each room and pairs of students were assigned to their first prompt. Students completed these activities in pairs, and were assigned a role of either speaker or listener. Speakers were instructed to explain their position in as much detail as possible, and listeners were instructed to ask questions if speakers could not continue speaking. After the timer rang, I asked the students to move in a clockwise direction to the next prompt. Each student completed two prompts as a speaker, and two prompts as a listener, that is, roles were switched each time. The students did not change partners during these activities. In classes where there were an uneven number of students, one student paired with the instructor.

The Discussion 2 activity enabled students to expand upon their opinions related to the ideas covered in the discussion preparation activities. A full list of the

Discussion 2 questions can be seen in Appendix E. For the extended discussions, the students were assigned into groups of four members, or three members if there were absent students, and whenever possible, students were placed into groups that did not contain their partner for the discussion preparation activity. In cases where there were only five students present, the discussion was conducted in one group of five students. Students participated in the discussions sitting around a table with a printed copy Discussion 2 questions from Appendix E available to them. Time was allotted to these discussions on a four-minute per participant basis, that is, for groups of three students, the discussion lasted for 12 minutes, while for groups of five students the discussion lasted for 20 minutes. The time was measured using an electronic timer. During the discussions, there was no teacher intervention and students had to continue speaking for the allotted time. The students were familiar with this policy as it was identical to the first semester discussion course that they had all completed.

The 3/2/1 Tasks

The 3/2/1 speaking task prompt was the same in each lesson and simply asked students to repeat the ideas they had shared during the class. The full prompt can be seen in Appendix F. The 3/2/1 tasks used in this study were taken from Lessons 3, 4, 6, 7, 8, 10, 11, and 12. Due to both time restrictions and the relatively low level of students' productive skills, the 4/3/2 activity used in previous research (Arevart & Nation, 1991; De Jong & Perfetti, 2011; Maurice, 1983; Nation, 1989) was shortened to 3/2/1. These activities were used to measure short-term fluency, as gains in fluency in successive rounds can be due to priming of word meanings and forms rather than evidence of long-term fluency development (De Jong & Perfetti, 2011). This activity was conducted after the final discussion in all regular lessons. In discussion test lessons, the task was completed in the first 45 minutes of class with slightly different wording; therefore, the performances from these activities were not used in this study. Students were paired with another student from a different discussion group than their own, unless there was only one group of five students. In cases where there were an uneven number of students, I joined as a listener for the first cycle of speeches and two students listened to one speaker when students switched roles. To avoid planning effects, where possible, students alternated as first speaker or listener each week.

Monologue Speaking Tasks

The monologue speaking tasks were two-minute talks on a topic not addressed in class. As the topics were novel to students, gains across the tests could arguably be seen as evidence of long-term fluency development and proceduralization, as students have not been primed prior to speaking (De Jong & Perfetti, 2011). The topics can be seen in Table 5, and a full list of the questions can be found in Appendix G. The monologue speaking tasks were recorded in the first 30 minutes of Lessons 4, 8, 12, and 14. A practice monologue test was also conducted in the first lesson. These lessons were selected for the following reasons: Lessons 4, 8, and 12 were review lessons where no new language was introduced, and Lesson 14 was the wrap-up lesson in which teachers could introduce activities of their own choosing. These tasks were conducted in the first third of the class to reduce any effects of priming that could influence student performance. Furthermore, the topics of these tasks had not been discussed in class to minimize any repetition effects. In an attempt to make the tasks as similar as possible, each test contained a question with a ‘Do you prefer *X* or *Y*? Why?’ structure. This structure was also common in many of the textbook activities; therefore, the students were familiar with answering these kinds of questions. The students participated in these activities in pairs, acting as both a speaker and a listener. Listeners were instructed not to ask questions or interrupt speakers. In cases where there were an uneven number of students, I joined as a listener for the first cycle of speeches and two students shared one speaker when the students switched roles. Before each round of speeches, one minute of planning time was given to speakers; however, they were instructed not to take notes. To avoid priming effects, speakers who spoke in the second round answered a different question than the student that they had listened to. The test questions were administered in a different order for each class, as can be seen in Table 5.

Table 5. *Administration of Monologue Speaking Tasks*

Week	Class	First Speaking Task	Second Speaking Task
4	A	3: YouTube & movies	4: TV & games
	B	5: Sport	6: Music
	C	7: Seasons	8: Celebrations
	D	9: Eating	10: Free time
8	A	5: Sport	6: Music
	B	7: Seasons	8: Celebrations
	C	9: Eating	10: Free time
	D	1: Travel	2: Where to live
12	A	7: Seasons	8: Celebrations
	B	9: Eating	10: Free time
	C	1: Travel	2: Where to live
	D	3: YouTube & movies	4: TV & games
14	A	9: Eating	10: Free time
	B	1: Travel	2: Where to live
	C	3: YouTube & movies	4: TV & games
	D	5: Sport	6: Music

First Language Fluency

First language fluency was measured during an in-class activity during Lesson 12. Students who were absent from this class completed the activity in either Lesson 13 or 14. Following the monologue test, the students were given a set of different questions (see Appendix H) and asked to answer the question in Japanese. Students were instructed that the purpose of collecting a Japanese speech sample was to compare it with the speed of their English speech, and asked to speak at their regular speaking speed. As no Japanese was spoken in this class, prior to the activity, the students were told to ask each other ‘How was your week?’ in Japanese’ for 2 minutes in an attempt to reduce unnecessary distractions from switching languages. Following this task, the students spoke into an IC recorder while addressing a listener in the same fashion that the English monologue speaking tests were conducted. In the same fashion as the monologue tests, for the classes that had an odd number of students, the teacher, who had an intermediate level of Japanese, acted as a listener for one student. This student then joined another pair as a second listener in the second delivery.

The Extraversion/Introversion Questionnaire

An extraversion questionnaire was administered to students during Lesson 12 of the discussion course. As mentioned in the previous chapter, this construct has been hypothesized to benefit second language learning (Dörnyei & Ryan, 2015), although there have been some contradictory findings in the small number of studies that have been conducted (Dewaele & Furnham, 2000; Van Daele et al., 2006). This questionnaire was completed during class time in the remaining five minutes of the lesson. Three students were absent from this lesson and did not complete the questionnaire. The items in this questionnaire were taken from a questionnaire that had been previously used with Japanese university students and shown good validity (Apple, 2011). For the study, I used the Japanese translation that had been developed by Apple (2011).

The eight extraversion/introversion items (see Appendices I and J) were taken from a previous study (Apple, 2011) of Japanese university students and personality factors. Following this study, students marked their responses on a 6-point Likert scale. The poles of the scale were marked (1 = *Strongly disagree,* 6 = *Strongly agree*); points 2 through 5 were unlabeled.

The Willingness to Communicate Questionnaire

A WTC questionnaire was administered to students during Lesson 12 of the discussion course. Although there have been no studies investigating the link between WTC and CAF factors, there has been one study which has shown WTC to predict student use of the L2 in the classroom (Munezane, 2014), and research has also shown WTC to correlate strongly with extraversion (Oz, 2014). This questionnaire was completed during class time in the remaining five minutes of the lesson. Three students were absent from this lesson and did not complete this questionnaire. The items were taken from a previous study that had been used with students at the university in which this study took place (Doe, 2013), which had shown acceptable levels of dimensionality and had been able to distinguish student levels. As the purpose of the questionnaire was to measure willingness to communicate in the discussion class itself, all the items were based around behaviors that students carried out during the discussion (see Appendices K and L for a list of the items and their Japanese translations). When constructing the items, I attempted to include a range of behaviors that would be relatively easy and difficult for students to be willing to do in class. Behaviors that were hypothesized to be easy were routine parts of a discussion such as giving reasons and talking about preferences. Behaviors that were hypothesized to be relatively difficult for students to be willing to do were relatively uncommon events such as disagreeing with other speakers and being the first person to speak in a discussion. Following Weaver (2010), students marked their responses on a 4-point Likert scale (1 = *Definitely not willing,* 2 = *Probably not willing,* 3 = *Probably willing,* 4 = *Definitely willing*).

Procedures

In this section, the procedures used to collect and record data are described. Student participation in all of the activities was transcribed for the duration of each activity. Speech that was produced before the activity began or after the activity finished was not included. This policy was implemented to ensure that all students were recorded for approximately equal speaking time for each activity. As the students had completed the first semester of the discussion class and were enculturated into the program, they rarely went off topic or stopped speaking during discussions and activities. Furthermore, as the classes had an average of eight students per group, they were aware that I could easily monitor their performance, which contributed to their course grades.

All of the tasks used in this study were recorded with six Sony IC BX-332 recorders. To accustom the students to their presence, all activities in every class were recorded, including those not used in this study. An exception to this was the test lessons, where no data were recorded. Each recorder was turned on at the beginning of each class and the full 90-minute lesson was recorded.

All paired activities—discussion preparation activities, the 3/2/1 tasks, and the monologue tests—were recorded with an IC recorder held by the speaker, with each pair standing in the corner of a large classroom to minimize potential interference from other speakers. I decided the pairings of students before each lesson and made adjustments if students were absent. Pairings were arranged so that the students spoke to each member of the class an equal number of times whenever possible. After finishing a speaking turn, the listeners moved to a new speaker, whilst the speakers remained holding their recorder in the same position. After finishing their speaking turn, the listeners and speakers switched roles. The speakers handed the recorder to the listener and remained in the same position to act as the first listener for the speaker. In cases where there were an uneven number of students in class, I joined as a listener for the first cycle of speeches and two students listened to one speaker when students switched roles. During pair work activities, the students did not have access to materials other than the questions in Appendices B and D. Dictionaries, textbooks, and other written notes were not permitted. All paired activities were timed using an accurate timer with a loud bell to signal students when to stop speaking.

Interactive discussions of three to five members were recorded with an IC recorder placed in the center of the table, around which students were seated, with each table situated in the diagonally opposite corner of a large classroom to minimize potential interference from the other group. These recorders were placed on the table at the beginning of the 90-minute lesson and the entire class was recorded. I decided the groupings for each discussion before each lesson and made adjustments if students were absent. Discussion groups were arranged so that the students spoke to each member of the class an equal number of times. Whenever possible, the discussions consisted of an equal gender balance, that is, two male and two female speakers. In cases where there were a different number of students in each group, i.e. three speakers in one group and four in the other, the discussions lasted for the full 16 minutes, however only the first 12 minutes were analyzed for the smaller group of three, to maintain four minutes of speaking time for each student in the data analysis. There was no teacher involvement or intervention in the discussion activities. During the interactive discussions, the students did not have access to materials other than the questions in Appendix C. Dictionaries, textbooks, and other written notes were not permitted. All paired activities were timed using an accurate timer with a loud bell to signal students when to stop speaking.

All recorded data were saved to a computer WAV file. These files were then separated and saved into separate files for each speaking activity using Audacity software. All files were labeled according to activity type, date of recording, and the name of participants. To ensure confidentially, pseudonyms were assigned to each student. These files were then transcribed into CHAT format compatible with CLAN software (MacWhinney, 2000). The large size of the classrooms in which the study was conducted meant that there was very little interference from other groups during the recordings. For sample transcripts of the monologue tests, see Appendix M, and for the 3/2/1 speaking activities see Appendix N.

The Design of this Study

The data collection schedule is detailed in Table 6. As the course was following a set curriculum, all students followed the same order of topics.

Table 6. *Data Collection Schedule*

Lesson	Topic	Collected data
1	Introduction	
3	Language 2	Discussion 1: Studying Abroad 3/2/1 Task 1
4	Fashion 1	Monologue Test 1 Discussion 2: Young People and Fashion 3/2/1 Task 2
5	Fashion 2 (Discussion Test 1)	
6	Media 1	Discussion 3: Is New Media Good for Us? 3/2/1 Task 3
7	Media 2	Discussion 4: Do Celebrities Make Good Role Models? 3/2/1 Task 4
8	Globalization 1	Monologue Test 2 Discussion 5: The Globalization of Japanese Culture 3/2/1 Task 5
9	Globalization 2 (Discussion Test 2)	
10	Human Rights 1	Discussion 6: Public Manners, Public Rules 3/2/1 Task 6
11	Human Rights 2	Discussion 7: The Right to Life 3/2/1 Task 7
12	Gender 1	Discussion 8: Gender Stereotypes Monologue Test 3 3/2/1 Task 8 Extraversion and WTC Questionnaire L1 Fluency Measure
13	Gender 2	
14	Wrap-up	Monologue Test 4

Data Coding Procedures

The recorded data from the tests and activities were then transcribed, entered into Microsoft Word, and analyzed into CAF and other related measures. One noticeable feature of the monologue speaking tests was that the students began speaking at different times, probably due to the fact that they had to generate content for their speeches, and it being the first speaking activity of the lesson. Furthermore, as there were common distractions, for instance students encouraging each other at the beginning of the recording, it was decided to reduce the analysis to the first 90 seconds of these recordings, beginning from the time the speaker started speaking. This decision was an attempt to reduce interference that could have resulted from these circumstances, as the differences in starting times could have obscured the utterance fluency measures.

AS units were chosen for the principle means of dividing the transcribed data into linguistic units. This unit has been defined as "a single speaker's utterance consisting of *an independent clause or sub-clausal unit*, together with any *subordinate clauses* (s) associated with either" (Foster, Tonkyn, & Wigglesworth) 2000, p. 365, italics in the original). The AS unit was chosen for several reasons. First, it has been increasingly used in SLA research concerned with the analysis of L2 speech, enabling results to be more easily compared among studies. Second, Foster et al. provided relatively clear instruction on how to analyze text into AS units, and they listed guidelines to deal with phenomena in spoken language that alternative measures, which have been generally been developed to analyze written texts (e.g., t-units), do not account for. This latter point is important given the many differences between written and spoken language. The definitions of the clauses considered AS units are:

- At a minimum, independent clauses must contain a finite verb (e.g., ‘Many students think they need to be fashionable.’)
- Independent sub-clausal units are either (a) phrases that contain ellipses but can be understood by the context of the discussion (e.g., ‘Twenty thousand yen’ in response to the question ‘How much do you spend on fashion every month?’) or (b) irregular or non-sentences (e.g., ‘very bad shop.’)
- At a minimum, subordinate clauses must contain “a finite or non-finite Verb element plus at least one other clause element (Subject, Object, Complement, or Adverbial)” (Foster et al., 2000, p. 366). For instance, ‘I did not have the opportunity :: to visit Japan’ would be counted as two clauses, while ‘I did not have the opportunity to visit’ would be counted as one. The symbol “::” is used to mark clause boundaries.
- If connected phrases have coordinated verbs, the transcriber must decide whether they are part of the same or different AS units. The criteria for this decision is that if the first phrase ends with rising or falling intonation and is followed by a pause of 0.5 seconds or longer, it is considered to be a separate AS unit. For example, ‘I feel a lot of pressure about what to wear (0.7) and how to decide what I should wear to school’ would be counted as two AS units provided there was rising or falling intonation in the first unit.
- Within an AS unit, subordinate clauses can take adverbial functions provided that it is “within the same tone unit as a least one of the other preceding clause elements

of the AS-unit" (Foster et al., 2000, p. 367). For example 'I went to shopping yesterday at Sunshine City' would be counted as one AS unit provided it meets these criteria.

The criteria for dealing with breakdown phenomena (false starts, repetitions, and self-corrections) common to spoken language are:

- False starts are marked in brackets as such and included in the AS unit in which they occur. For instance, '{I want no} university students need to sometimes be fashionable.'
- While repetitions are largely dealt with in the same manner as false starts, for example, '{I want to go} I want to go shopping for clothes with you' is included in the AS unit in which it occurs, repetition used for rhetorical effect. For instance, 'it's very very complicated' are not bracketed and included in the AS unit.
- Self-corrections are also bracketed and included in the AS unit in which they occur. An example of this is, 'we often {thinks} think about what our friends think'

Foster et al. (2000) also noted the need for researchers to decide upon a principled manner of whether to include the linguistic information included in false starts, repetitions, and self-corrections in their analyses, however as they mention, many researchers do not disclose such information in published studies. Following De Jong and Perfetti (2011), the fluency measures in this study included the students' false starts, repetitions, and self-corrections. One reason for including these phenomena is that they are often used by native speakers to signal a variety of functions, for example, to inform listeners that there will be some delay in speech, which begs the question of whether such phenomena interfere with or facilitate smooth communication (De Jong, Steinel, Florijin, Schoonen, & Hulstijn, 2012b). However, for the accuracy and complexity measures described below, these phenomena were excluded from the analyses. This decision was made for the accuracy measures because many false starts and reformulations were attempts by the speaker to either prevent or correct errors. These phenomena were also excluded from the complexity measures as the repeated words would artificially reduce the level of lexical complexity in the vocabulary analyses. Similar criteria were used by De Jong and Perfetti (2011), who included these phenomena in their fluency measures, but excluded them from their lexical complexity measures although their study did not look at complexity and accuracy.

A final feature common to spoken language is topicalized noun phrases occurring within clauses, for example, 'the clothes shops in Ikebukuro they are so expensive.' If an analyst were to use alternative measures such as t-units, it is unclear how such phrases should be measured. In a similar manner to coordinated verb phrases, Foster et al. (2000) advised including topicalized phrases in the AS unit in which they appear. The one exception to this rule is when such phrases are marked by falling intonation and followed by a pause of 0.5 seconds or longer. This guideline was followed in this study.

Complexity

The levels of syntactic and lexical complexity in the monologue tests were determined by two measures. Syntactic complexity was measured by the total number of clauses divided by the total number of AS units and the number of words per AS-unit. While the total number of clauses divided by the total number of AS units is a commonly used measure, it can fail to distinguish variation in complexity for beginning level learners who rarely use subordinate constructions (Ellis & Barkhuizen, 2005, p. 155). Therefore, the number words per AS-unit was added as a supplementary measure. For lexical diversity, the MTLD (measure of textual lexical diversity) was used as an alternative to type-token ratio, as the latter does not take text length into account. While this measure produces results that are less dependent on text length, it can still be affected by this factor (Jarvis, 2013); however, a preliminary analysis showed that there were no significant correlations between the total number of syllables uttered and the MTLD in each monologue speaking test. This measure was calculated using an online analysis tool (http://cohmetrix.com/), the use and function of which is outlined in McNamara, Graesser, McCarthy, and Cai (2014). As mentioned above, false starts, repetitions, and self-corrections were excluded from all complexity analyses. L2 words that were not listed in an English dictionary were also excluded from the complexity analyses, however, all proper nouns from English and foreign languages were included.

Accuracy

The levels of spoken accuracy in the monologue tests were determined by three measures, the total number of errors per speech, the number of words per error, and the ratio of error-free clauses. These measures were selected because they have been shown to measure short-term changes in previous studies of oral language development (Delany, 2012; Skehan & Foster, 2012; Tonkyn, 2012). As mentioned above, false starts, repetitions, and self-corrections were excluded from all accuracy analyses in order to avoid artificially inflating the number of errors. Grammatical errors were also classified into four categories: (a) noun phrase errors (related to incorrect use of number/case, or determiners), (b) verb phrase errors (related to incorrect inflection or auxiliary verb use, (c) syntactical errors related to incorrect word order or the omission or misuse of essential clause elements, and (d) lexical errors (related to the incorrect use of an open class word, i.e. a lexical verb, noun, or adjective, or adverb), following Tonkyn (2012), who was able to discern growth in levels of accuracy with ESL students in a 10-week study by calculating the number of words per error type. This coding was carried out in case further analyses of the types of errors that occurred were necessary. L2 words that were not listed in an English dictionary were coded as lexical errors. An example is ‘I think it is {*mottanai* [wasteful]}. All proper nouns from English and foreign languages were included in the accuracy analyses.

Fluency

Speech from the 3/2/1 speaking task to measure short-term fluency development, and the monologue speaking tests to measure long-term fluency development was analyzed to determine the degree of change in oral fluency. Four measures of temporal fluency were included in these analyses, following De Jong and Perfetti’s (2011) fluency training research: the articulation rate, the mean length of pauses, the phonation/time ratio, and the mean length of fluent runs. Although including a large number of fluency variables reduces the statistical power of the analyses, there are several important reasons for including more than one measure.

First, research into the temporal variables theorized to form fluency "has not revealed compelling, consistent patterns of oral production features that might be considered reliable markers of fluency" (Segalowitz, 2010, p. 4), which suggests that researchers should consider more than one fluency variable. Second, several researchers have argued that fluency development is the result of change in a combination of measures. Towell et al. (1996) argued that fluency development is shown by an increase in the length and complexity of utterances between pauses, i.e., an increase in the mean length of fluent run while pause lengths remain stable. De Jong and Perfetti (2011) also found that some speakers made gains in fluency with decreasing pause lengths, an increasing phonation/time ratio, and a stable mean length of fluent run, and also argued that this was a sign of fluency development.

To calculate the articulation rate, the number of syllables per minute of speech was calculated. Japanese words and nonverbal fillers in either English or Japanese were not counted as syllables. Syllables were counted by pasting the transcripts into an online syllable counter (http://www.syllablecount.com/). Additionally, six minutes of data taken from three 2-minute monologue tests from different speakers was counted and compared with the computed results. There were no disagreements between the two counts. Japanese words included in the analyses, such as place names, were checked, and if necessary, adjusted. Each Japanese mora was counted as one syllable.

To calculate the mean length of pauses, the recordings of student data were transcribed using PRAAT 5.3.53 software (Boersma, 2011). Pause boundaries were measured by listening to the recording and viewing the graphic waveform; 250 ms was decided as the minimum length of pause. De Jong and Perfetti (2011) noted that there has been a lack of standardization when deciding the cut-off point for pausing, for example, Kormos and Denes (2004) used 250 ms, while Towell et al. (1994) used 280 ms as the minimum length of pause. While De Jong and Perfetti argued that the lower measure of 200 ms is preferable due to Lennon's (1990) findings, in which judges rated the majority of pauses 200 ms or longer as sounding dysfluent, follow-up research revealed that cut-off points in the range of 250-300 ms had the strongest correlation with L2 proficiency (De Jong & Bosker, 2013). Therefore, the lowest rate in this range, 250 ms, was chosen as the cut-off threshold for measuring pauses.

Previous researchers have also trimmed pauses in an attempt to minimize the influence of the environment. In De Jong and Perfetti's 2011 study, an upper limit of 2.5 standard deviations above the mean length of pause in a specific speech was set. These pauses were trimmed in order to eliminate pauses that might not have been the result of a student's fluency, as they might have been distracted or reacting to some occurrence in the classroom. However, it was my impression that the majority of long pauses were the result of students' attempts to generate content for their speeches rather than caused by distractions, as many long pauses occurred in the middle of clauses, and seemed to be the result of students taking time to think about the vocabulary and/or grammar needed to complete their utterance. Therefore, the length of pauses was not adjusted. Finally, nonverbal fillers and hesitation phenomena were not counted as syllables and included in pauses, following De Jong and Perfetti.

For the monologue speaking tests, pauses were further separated into two categories: mid-AS unit and end-AS unit pauses. Research comparing the placement of pauses between native and L2 speakers has shown that a significant difference between these two groups is that L2 speakers tend to pause more mid clause (Riazantseva, 2001; Tavakoli, 2011). Tavakoli argued that although difficult to specify, the reason for this is probably due to cognitive factors associated with "information processing, repair mechanisms, monitoring, and online planning" (p. 77). Recent researchers have also argued that successive repetitions are needed to reduce mid-clause pauses, which are the result of learners needing to devote mental attention to matters related to lexis and syntax (Lambert, Kormos, & Minn, 2016). For these reasons it was decided to compare the amount of mid- and end-AS unit pauses for the monologue speaking tests, as reductions in mid-AS unit pauses suggest that the learners were becoming more able to fluently control their language.

The phonation/time ratio was determined by dividing the total speech time actually spent speaking, i.e. with all pauses deducted, by the total time of the speech. The mean length of fluent run was determined by calculating the mean number of syllables produced in utterances between pauses longer than the cut-off rate of 200 ms.

Word Counts and Lexical Overlap

Word counts and lexical overlap were calculated for the discussion preparation activities, the discussion activities, and the 3/2/1 speaking tasks. The FREQ function of CLAN software was used to count the number of types and tokens used by each student in each activity. Following Arevart and Nation (1991), contracted words (e.g., can't, shouldn't) were counted as two words. The type lists were then pasted into an Excel spreadsheet, and pivot tables were created to determine the amount of lexical overlap in all five activities, in addition to those repeated solely in the 3/2/1 speaking tasks. See Verspoor, Lowie, van Geert, van Dijk, and Schmid (2011) for detailed instructions on how to create a pivot table of item frequency across a variety of texts.

Use of Directly Taught Formulaic Language

The use of directly taught formulaic language was calculated for the discussion preparation activities, the discussion activities, and the 3/2/1 speaking tasks. Formulaic language taught in the semester preceding the study was also included. For a complete list of the formulaic language directly taught in the course, see Appendix A. The COMBO function of CLAN was used to determine the number of times each student used directly taught formulaic language in each activity. Slight deviations from the phrasing, for example, *A different example is…* instead of *Another example is…* were counted as instances of formulaic language use. Inaccurate use of formulaic language, such as *A different example are…* were not counted as instances of formulaic language use. Formulaic phrases performing the same communication function with different wording were not included, for example *Let me illustrate my point by explaining…* were not counted as an instance of use of the formulaic utterance.

First Language Fluency

The L1 monologue speeches were transcribed into the Japanese syllabary *hiragana* by a native speaker of Japanese. In general, each character in hiragana is one mora; therefore, the total number of characters for each speaker was counted using the word count function on Microsoft Word. This total was then divided by 2 to determine the number of mora spoken per minute. This figure was used as the measure of first language fluency.

Expert Ratings of Perceived Fluency

The rating scale (see Appendix O) for perceived fluency was adapted from Nitta and Nakatsuhara (2014). This study was chosen because it was carried out with Japanese university students in a similar context; however, these students were slightly older, and as they were English majors, of a slightly higher proficiency level. The major changes to the scales were a reduction in the length of each outcome space descriptor, and a reduction in the number of outcome spaces, from eight to five. These changes were made as research has shown that scales with more than six outcome spaces make too many demands on working memory (Nemoto & Beglar, 2014). Five outcome spaces were chosen as this was the number of outcome spaces in Nitta and Nakatsuhara that had descriptors. The other spaces were included as midway points between these descriptors.

The ratings were carried out by six raters, all of whom taught similar speaking classes to students at another institution. All of the teachers had a Master's degree in Applied Linguistics or a related field, and all of them had to rate their own students' oral fluency in the speaking courses that they taught. Approximately 20 minutes of rater training was provided, where the group was introduced to the rating scale and descriptors, and then asked to independently rate the recordings of four speech samples. The instructions given to the raters can be seen in Appendix O. The speeches on these recordings were very near to the average of each class' mean length of run.

After independently rating the samples, the raters discussed the scores that they had given each speaker and their rationale for doing so. None of the ratings in the training session differed by more than 1 point. Following from this, the raters were played 16 recordings of different monologue tests, two speakers from each class. These recordings that were selected were those where the range between the highest and lowest mean length of run was the largest. The raters were not informed that they would be listening to the same speakers twice. The mean length of run was selected as the variable of interest because it has been used as an indirect indicator of improvements in speaking automaticity (De Jong & Perfetti, 2011; Towell et al., 1996). The size of the range was considered important, as some of the differences appeared relatively small, and were often less than one syllable. The raters were played all 16 recordings in a randomized order, and each recording was rated without any discussion or consultation with each other.

Analyses

To answer the first research question, which looked at long-term fluency development, a two level multi-level model (MLM) analysis was conducted for each of the five fluency measures. For those variables which showed significant development over the study, follow-up analyses were conducted with the in-class performance measures and individual difference measures as independent variables. Similarly, the second research question, concerning short-term fluency development, was investigated with three level MLM analyses for four of the five fluency measure. Due to the large amount of data in these analyses, time constraints prevented an investigation of the mid-clause pause to end-clause pause ratio fluency measure. For those variables that showed significant development over the three deliveries of the 3/2/1 speaking task, follow up analyses were conducted with the in-class performance and individual difference measures as independent variables.

To answer the third research question, which concerned the relationship between the complexity, accuracy, and fluency variables, correlation analyses were conducted between the complexity, accuracy, and fluency measures for each of the monologue speaking tests. Finally, to answer the fourth research question, which concerned expert rater judgements of speaking fluency, correlation analyses were conducted between the average rating given on each performance with each of the five fluency measures.

Multi-Level Modeling

MLM analysis was used to measure fluency development and the variables that could influence its trajectory. This statistical technique can be used when researchers wish to investigate the degree to which development in variables at lower levels could be influenced by those at higher levels. For example, a student's development in a skill can be influenced by factors such as their overall motivation for the class, or their teacher's level of training and expertise. Furthermore, MLM is capable of dealing with several challenges that researchers wishing to analyze repeated measures of language development in classroom settings face. First, MLM is able to account for dependence among observations, meaning that the assumption of independent measurements required in general linear modeling can be disregarded (Field, 2009). This is particularly relevant for classroom-based research where an individual student's performance can be influenced by a variety of environmental factors and issues arising from subjects in one group being more similar to each other than to subjects in other groups (O'Connell & McCoach, 2008). Second, MLM can also deal with missing data (Goldstein, 2011), which is ideal for classroom-based research where students can be unexpectedly absent from a class during the data collection period. All multi-level models were conducted on HLM software, version 7.01 (Raudenbush, Bryk, & Congdon, 2013).

MLM models were used in this study in order to investigate change in both short- and long-term fluency development. Four models were run for both fluency types, one model for each measure of fluency. In these models, the dependent variables were the mean length of run, the mean length of pause, the articulation rate, and the phonation/time ratio. These measures were selected because several researchers have argued that change in one fluency measure might be offset by increases in others, therefore more than one fluency measure must be examined (De Jong & Perfetti, 2011; Towell et al., 1996).

For long-term fluency development, four two-level models were used to determine whether utterance fluency measures developed over the semester. These measurements were taken from each of the four monologue speaking tests, and were the dependent variables situated at the first level of the model. A variety of measures based on performance in class activities, such as the total number of words spoken throughout the semester, and individual differences, such as the extraversion and WTC measures, were added to the second level to determine whether they covaried with changes in the dependent variables over the course of the semester.

The first step in undertaking these MLM analyses was to run an unconditional model to determine whether the dependent variables—mean length of run and the mean length of pause—made any significant changes over the semester. In this model, the only predictor variable was the time when the test was conducted. This is represented by the equation:

$$Y_{ti} = \beta_{0ti} + \beta_{1t}(TIME) + r_{ti}$$

Y_{tij} is the mean length of run or mean length of pause at time t for learner i. $\beta 0_{ti}$ represents the fluency variables on the first test, and $\beta 1_t(TIME)$ represents the growth rate for learners across the semester. Finally, r_{ti} represented the residuals associated at time t with learner i.

The second level of the model contained variables related to the students' language output over the semester, in additional to individual difference variables. This allowed for analyses to be conducted to determine whether any of these variables could predict growth. In the first level of the equation below, β_{0i} represents the students' initial fluency on the test, while β_{1i} represents the growth across the four tests. The performance and individual difference variables can then be entered into the second level to investigate the degree to which they can predict development in utterance fluency.

$$\beta_0 = \gamma_{00} + u_0$$

$$\beta_1 = \gamma_{10} + u_1$$

For example, this equation investigates the relationship between the students' levels of extraversion and their fluency development:

$$\beta_{1i} = \gamma_{10} + \gamma_{10}\,(\mathrm{EXTRAVERSION}) + u_{1i}$$

For short-term fluency development, four three-level models were used to investigate fluency development. These fluency measures were calculated for each speech in the 3/2/1 activity, and were at the lowest level in the model. The 3/2/1 data were nested inside the class in which the 3/2/1 activity was conducted, and included in the second level were performance variables for that class: the average amount of repeated words, the average amount of words spoken, and the average amount of formulaic phrases used in the lesson where the 3/2/1 activity was conducted. This data were nested inside individuals at the third level, where students' levels of extraversion, willingness to communicate, second language proficiency, and first language fluency were included.

Similar to the long-term models, the first step in undertaking these MLM analyses was to run an unconditional model to determine whether the dependent variables made any significant changes over the delivery of the three speeches in the 3/2/1 activity. In this model, the only predictor variable was the order of the speech delivery. This is represented by the equation:

$$Y_{ti} = \pi_{0ti} + \pi_{1ti}(SPEECH) + e_{ti}$$

Y_{tij} is the mean length of run or mean length of pause at time t for learner i. π_{0ti} represents the fluency variables on the first speech, and $\pi1_{ti}(SPEECH)$ represents the growth rate for learners across the three deliveries. Finally, e_{ti} represented the residuals associated at time t with learner i.

The second level of the model contained variables related to the students' language output in the class during which the activity was conducted. This allowed for analyses to be conducted to determine whether any variables could predict growth. In the first level of the equation below, π_{0i} represents the students' initial fluency on the first delivery of the speech, while π_{1i} represents the growth across the three deliveries. The student performance variables can then be entered into the second level to investigate the degree to which they can predict development in fluency.

$$\pi_{0i} = \beta_{00} + r_{0i}$$

$$\pi_{1i} = \beta_{10} + r_{1i}$$

For example, this equation represents the relationship between the total amount of words repeated in all of the lesson's five activities and fluency development:

$$\pi_{1i} = \beta_{10} + \beta_{11}(\text{the total number of words repeated in all five activities}) + r_{1i}$$

The third level of the model contained contextual variables related to the students' individual differences such as extraversion, WTC, and level of English proficiency. Due to the students' limited exposure to English throughout the semester (4.5 hours per week), these variables were assumed to remain relatively stable. Extending the model to a third level allowed for analyses to be conducted to determine if any contextual variables covaried with fluency growth, for example, whether students who produced more language were more highly extraverted than their peers who did not, represented by these equations:

$$\beta_{00} = \gamma_{000} + u_{00}$$

$$\beta_{10} = \gamma_{100} + u_{10}$$

$$\beta_{11} = \gamma_{110} + \gamma_{111}(\text{EXTRAVERSION}) + u_{11}$$

These individual difference variables can also be entered at the other two levels in the model to determine whether they have an effect on initial levels of fluency or its growth trajectory.

The Rasch Model

Rasch model analyses were conducted to examine the validity of several measures used in this study: the extraversion and willingness to communicate questionnaires, and the rubric used by the expert raters to measure perceived fluency. There are several benefits to using the Rasch model to investigate the validity of the measures used in a study. First, the raw scores or ratings can be converted from ordinal to interval scales. In addition, fit indices derived from these analyses can be used to determine the degree to which the data from the persons and items fit the model, that is, the amount of variation between the actual data and the responses predicted by the Rasch model. Finally, the dimensionality of the measure used to represent a construct can also be determined through Rasch principal components analyses of residuals between expected and observed responses.

The most basic Rasch rating scale is the dichotomous model, which estimates the probability of a subject selecting or answering an item. First, person estimates are obtained by calculating the ratio of the percentage of items a person scored correctly to the percentage of items that were answered incorrectly, and taking the natural logarithm of this ratio. Similarly, item difficulty estimates are obtained by calculating the ratio of the number of persons who scored the item correctly with the number of those who did not. Once these values have been obtained, the probability of a person answering an item correctly can be determined from the following equation:

$$P_{ni} (x = 1) = f (B_n - D_i)$$

where $P_{ni} (x = 1)$ is the probability of person n answering item i with a score (x) of 1 is a function (f) of the person's ability (B_n) minus the difficulty of the item (D_i).

When applying the Rasch model to polytomous data, such as the Likert scale questionnaires and rubric used in this study, the rating scale model can be used to take the more complex responses into account. The major difference of this model is that it creates difficulty estimates for each threshold between the response categories on the items. Similarly to item measures in the dichotomous model, item thresholds are modeled as the probability of a person choosing one response category over another. This can be expressed as:

$$P_{ni1} (x = 1 / B_n, D_i, F_1)$$

where P_{ni1} is the probability of person n selecting the response category 1 instead of the first response category 0 on item i. Again B_n is the person's ability, D_i is the difficulty of the item, and F_1 represents the difficulty of the first threshold in the response category.

A more complex situation is when additional factors need to be taken into account when validating a rating scale. For example, the experts who rated the speech samples in this study may have varied in their severity. The many-facets Rasch model is able to control for factors such as these that might have a systematic influence on the data. This model is able to calibrate difficulty estimates for each facet in the model, such as the severity (C) of expert rater j. When calculating the probability estimate of person *n* responding to item *i* for a category threshold *k*, the severity of rater *j* is taken into account. This can be expressed through the following equation:

$$P_{nik} = \frac{e^{(B_n - D_i - F_k - C_i)}}{1 + e^{(B_n - D_i - F_k - C_i)}}$$

where these estimates are a function of the ability person B_n, the difficulty of item D_i, the difficulty of the threshold F_k, and the severity of the rater C_j.

Rasch fit indices. Given the complexity involved in measurement, it would be highly unlikely that the estimates produced by the Rasch model would match the observed data with 100 percent agreement. However, an important part of validating a scale is to the identify persons and items that behave in unpredictable ways. Therefore, fit indices can be examined to determine whether there are large degrees of deviation between the data and Rasch estimates. The mean square outfit statistics are unweighted averages of the residual variance (Z_{ni}) across persons and items, and can be expressed as:

$$\text{outfit} = \frac{\sum Z_{ni}^2}{N}$$

In contrast, mean square infit statistics are weighted residuals which place more importance to the unexpected responses that occur near a person's ability estimate, or an item's difficulty estimate. The infit statistics therefore, are weighted averages of the residual variance (W_{ni}) across the persons and items and can be expressed as:

$$\text{infit} = \frac{\sum Z_{ni}^2 W_{ni}}{\sum W_{ni}}$$

Fit statistics can also be standardized into *t* or *Z* distributions. Bond and Fox (2015) note that the mean square fit statistics can be examined to determine how much misfit occurs between the data and the estimated model. On the other hand, the standardized values can be examined to determine how likely the misfit between the data and the estimated model is.

Reliability and separation indices. Another use of Rasch model analysis is to determine how reliably persons and items are separated by the scale used to measure the object of interest. The person estimates are used to determine the degree to which the scale can reliably separate the persons who answered the items. This estimate can be calculated using the following formula:

$$R_p = \frac{SA_p^2}{SD_p^2}$$

where the amount of variance that is able to be reproduced by the Rasch model (SA_p^2) is divided by the total person variability (SD_p^2), an in the same way as traditional measures of reliability such as Cronbach's alpha, the estimate's values range between 0 and 1 (Bond & Fox, 2015).

Another measure that can be used to determine the degree to which the scale is able to spread person scores on the measure is the person separation index (G_p). This is obtained by dividing the adjusted person standard deviation (SA_p) by the average measurement error (SE_p), and is expressed by:

$$G_p = \frac{SA_p}{SE_p}$$

Similarly, the item separation index can be obtained by substituting the item variance for person variance:

$$G_i = \frac{SA_i}{SE_i}$$

A final calculation can be carried out to determine the number of “measurably distinct strata” (Bond & Fox, 2015, p. 354) that the person or item data can be separated into. This can be determined by the following equation:

$$\text{Strata} = \frac{4G-1}{3}$$

Rating scale functioning. Rating scales need to be carefully examined to determine whether they are functioning in a meaningful way. Linacre (2002) outlines eight criteria that can be used to evaluate the effectiveness of a rating scale. First, each category should have a minimum of ten observations. Second, there should be regular observation distribution across the categories. Third, the observed average measures of the persons in the category should increase with the rating scale category. Fourth, the outfit mean-squares should be less than 2.0. Fifth, the Andrich thresholds should advance, that is, respondents who exhibit higher values of the construct should have higher probabilities of being observed in the higher categories of the rating scale. Sixth, the ratings should imply the measures, and the measures should imply the ratings. That is, there should be a relationship between the measures and the mean of the expected ratings. Seventh, the Andrich thresholds should advance by at least 1.4 logits, if the scale has three points. Elwood (2011) calculated the necessary threshold calibrations needed for larger scales and reported a minimum separation of .59 logits for a six category Likert scale such as those used in the questionnaires in this study. Finally, the Andrich thresholds should advance by less than 5.0 logits.

Rasch principal components analysis of item residuals. Rasch model analyses can also be used to determine the dimensionality of items that have been designed to measure a construct. The principal components analysis identifies the amount of variance that explains the most variance and then examines the remaining variance to determine if there are any additional dimensions. Linacre (2015) states that a construct can be considered to be unidimensional if the variance explained by the measures is greater than 50 percent, and the variance explained by the first contrast is less than 10 percent, or has an eigenvalue of less than 3.0.

Summary

In this chapter, I have outlined the methods used to carry out this study. The participants, course materials and the class in which the study was conducted were described. Following this, the procedures for collecting the data for the study were explained. The measures and instruments used to investigate student development were also described. Finally, the chapter concluded with a description of how the data would be analyzed.

When reviewing the literature into second language fluency development, it is noticeable that the sample sizes for all of the studies are far smaller than those usually seen in other SLA research. The main reason for this is that the process of transcribing and analyzing the data to determine fluency development is particularly time consuming. The methods and procedures outlined in this chapter took approximately 1,100 hours to complete with a large proportion of that time spent on measuring the length of pauses.

In the next chapter, I give a detailed explanation of the measures used in the study and explained how they were screened before conducting the main analyses.

CHAPTER 4

PRELIMINARY ANALYSES

In this chapter I discuss the measures used in the study in more detail, and outline how they were examined for reliability and validity. The CAF variables were checked by a second researcher who coded a small portion of the data. For the individual differences variables, the dimensionality and model fit of the extraversion and WTC measures were investigated using the Rasch measurement model (Rasch, 1960), using WINSTEPS version 3.91.0 (Linacre, 2013). In addition to dimensionality and model fit, the inter-rater reliability of the judges of students' perceived fluency was also checked using a multi-faceted Rasch model.

Reliability of CAF Measures

To determine whether the CAF measures could be reliably measured according to the guidelines in the preceding chapter, data from four of the students (12.5% of the total sample size) from the first monologue speaking test was calculated by a colleague engaged in similar research and familiar with the relevant literature, and then checked for reliability. The four students were randomly selected from each class.

Complexity

Percentage agreements were calculated for the classification of student output into AS-units and clauses, two measures for each of the four students. Initially, the percentage agreement was 87.5%. An analysis of the results indicated that the disagreements arose from the classification of gerunds. After consulting Foster et al. (2000), it was decided that gerunds should be counted as subordinate clauses, so long as they met the condition of being uttered with “at least one other clause element (Subject, Object, Complement or Adverbial)” (p. 366). Therefore, *shopping in Ikebukuro is my favorite thing to do after classes*, would be counted as one AS-unit with two clauses, whereas *shopping is my favorite thing to do after classes* would be counted as one AS-unit with one clause. The data were then re-checked and re-coded, and agreement was 100%.

Accuracy

Percentage agreements were calculated for the classification of student output into noun phrase, verb phrase, syntax, and lexical errors. Initially, the percentage agreement was 81.25% (13 of 16 errors). An analysis of the results indicated that the disagreements arose from the classification of three different errors. The first was related to the missing verbal element in the phrase *and the second reason in winter if we feel*, where one rater did not code the error based on the assumption that such an utterance would be acceptable in speech. After consulting Foster et al. (2000), it was decided this error should not be counted, as it could be seen as a topicalized noun phrase. The second discrepancy arose from coding the phrase *I can stay long*, where one rater coded the utterance as two noun phrase errors (a missing article and noun), and one rater coded the utterance as one noun phrase error and one syntax error. After consulting Tonkyn (2012), it was decided to code the utterance for one noun phrase error for the missing article, and one syntax error for the omission of a necessary phrase element, the missing noun. The final discrepancy arose from a similar issue regarding how to code subject verb agreement errors that followed case errors. Following Tonkyn's (2012) study of CAF development, both errors were counted, coded as a verb phrase error and a noun phrase error. The data were then re-checked and re-coded, and agreement was 100%.

Fluency

As the syllable and word counts were calculated using objective criteria, the only fluency data to be checked for reliability were the length of pauses. As student performances were obtained in a classroom, there was concern with the degree of accuracy that the duration of pauses could be measured because background noise or interlocutor comments can also appear on the wave image, making accurate measurement difficult. A comparison of the two analyses showed that the same number of pauses was counted by both analysts, however, there were many minor differences in the duration of the pauses. Table 7 shows the means and standard deviations for the mean length of pause in seconds, in addition to the differences between the two analysts' measurements. As can be seen, the measurements were similar with the largest difference between the analysts being 0.04 seconds. Due to the small size of these differences, the method used to measure pause duration was considered acceptable for the purposes of this study.

Table 7. *Reliability Analysis of Pause Durations*

	Rater 1: *M*	Rater 1: *SD*	Rater 2: *M*	Rater 2: *SD*	*M Difference*
Student 1	1.94	2.37	1.90	2.34	.04
Student 2	1.02	.92	1.04	.92	.02
Student 3	1.47	1.54	1.50	1.50	.03
Student 4	1.29	1.07	1.28	1.08	.01

Extraversion

The eight items on the extraversion questionnaire were entered into a Rasch rating scale analysis. The measure used a six-point Likert scale which was only marked at the extreme ends (1 = *Strongly disagree,* 6 = *Strongly disagree*); points 2 through 5 were unlabeled. The Likert-scale category functioning was examined with regard to Linacre's (2002) criteria outlined in the previous chapter, and can be seen in Table 8. The minimum of 10 observations per category was met, with the smallest number of observations being 19 in Category 6. There was a regular distribution of observations across categories, and the observed averages of the persons increased with the rating scale category. The outfit mean square statistic was below 2.00 for all categories. The Andrich thresholds advanced, and there was a relationship between the measures and the average expected ratings. Separation between categories was greater than the required .59 logits for a six-point scale (Elwood, 2011), and none were larger than 5.0. The category characteristic curves can also be seen in Figure 2, which demonstrates how the rating scale accords with the model expectations. After examining these results, it was decided that the Likert-scale was performing in a satisfactory manner.

Table 8. *Category Structure Functioning for Extraversion*

Category	Count	(%)	Infit MNSQ	Outfit MNSQ	Structure measure	Category measure
1 Strongly disagree	28	12	1.23	1.16	None	(-3.95)
2	39	17	1.05	1.07	-2.73	-2.06
3	41	48	.65	.69	-1.06	-.70
4	66	18	.96	1.08	-.37	.64
5	39	17	.58	.63	1.58	2.16
6 Strongly Agree	19	8	1.30	1.28	2.58	(3.89)

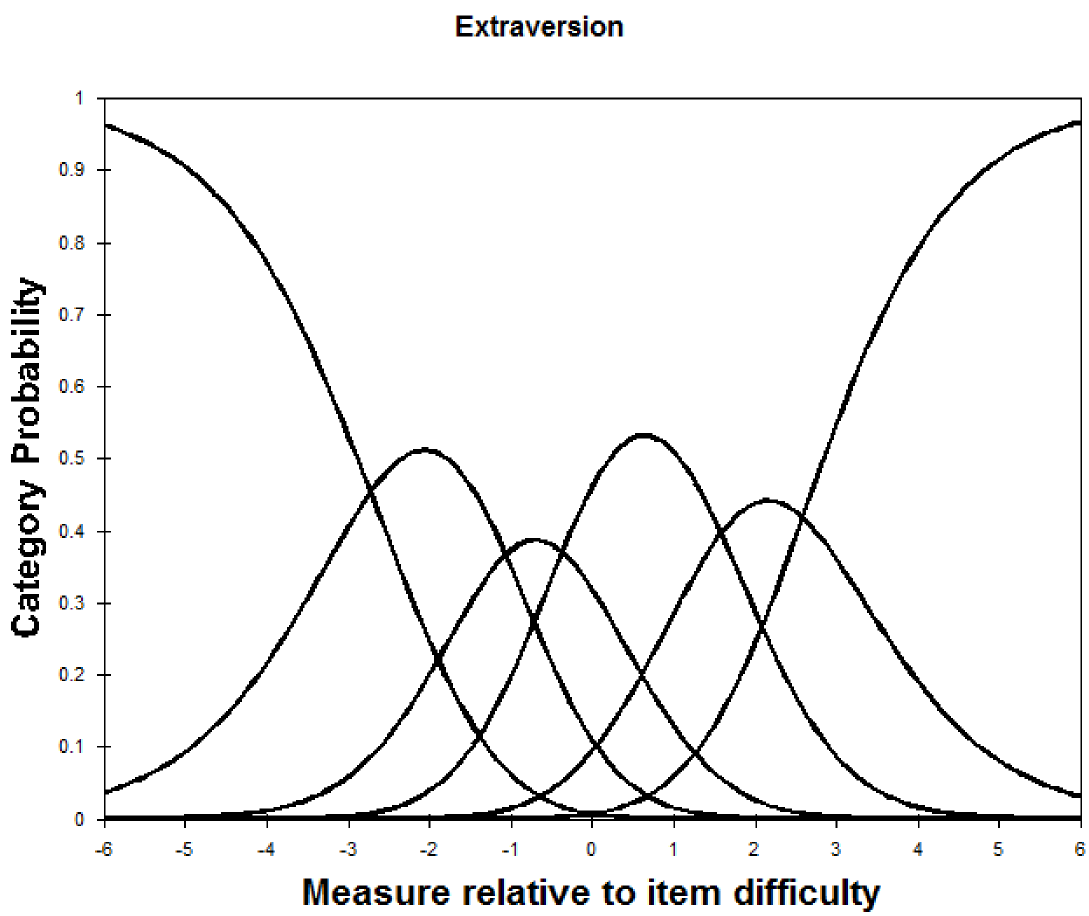

Figure 2. Rating scale performance of Extraversion measured on a 6-point Likert scale.

Table 9 shows the item fit statistics for the eight items, which shows that all of the items were within the acceptable Infit MNSQ value of 2.0 (Linacre, 2002). Item EXT5, *like to be the center of attention,* slightly underfit the model, although the

ZSTD Infit and Outfit were below 2.0, which indicates that the standardized misfit is not statistically significant. However, the person measures both with and without this item were exported to SPSS and a bivariate correlation analysis was run. The Pearson correlation was .99 ($p < .001$), which indicated that the removal of the item would do little change to the person measures. Based on this result, the item was included in the subsequent analyses. Items EXT3, *start conversations*, and EXT4, *like to draw attention to myself*, overfit the model. This overfit was probably the result of the lack of independence among the items, due to the similarity of items EXT3, EXT6, and EXT2, which all involved talking with people, and items EXT4 and EXT5, which involved being the focus of attention from others. Because overfit has minor implications for measurements—such as slightly increasing reliability coefficients (Bond & Fox, 2015)—these items were retained for the subsequent analyses.

Table 9. *Rasch Item Statistics for the Extraversion Items*

Item	Measure	*SE*	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt-measure correlation
EXT1	.71	.23	1.20	.79	.79	.79	.79
EXT2	-1.18	.24	1.14	.61	.61	.61	.83
EXT3	-.30	.23	.56	-1.95	-1.95	-1.95	.82
EXT4	-.46	.23	.46	-2.55	-2.55	-2.55	.82
EXT5	.82	.23	1.47	1.63	1.63	1.63	.79
EXT6	-.79	.23	.98	.01	.01	.01	.83
EXT7	1.09	.24	1.21	.81	.81	.81	.78
EXT8	.12	.23	.65	-1.49	-1.49	-1.49	.81

The Wright map (Figure 3) shows the students on the logit scale on the far left of the figure. The students are displayed as Xs according to their Rasch extraversion measures, with students with higher levels of extraversion toward the top of the map and persons with lower levels of extraversion toward the bottom of the map. The items are displayed on the right side of the figure according to their difficulty estimates, with the items that were more difficult to agree with at the top of the map and the items that were easier to agree with at the bottom. The average person measure was -.13, indicating that the items were well matched to the students. This is supported by the Wright map, which shows that people and items were fairly well distributed around the mean.

The extraversion measure had an item reliability of .89 and a person reliability of .92, item separation of 3.29, and person separation of 2.92. These results indicate that estimates were fairly well separated (Fisher, 2007). The Rasch PCA of item residuals analysis showed that 71.8% of the variance was explained by the Rasch model, surpassing Linacre's (2015) criterion of 50%. 25.8% of the variance (eigenvalue = 20.4) was explained by the items, and 9.3% of the variance was explained by the first residual contrast, with an eigenvalue of 2.64 (33% of the unexplained variance), meeting Linacre's (2015) eigenvalue criterion of under 3.0.

The items which were the most difficult to endorse were EXT7 *feel at ease in a large group*, EXT1 *am the life of the party,* and EXT5 *like to be the center of attention.* These items all relate to groups of people, which suggests that the students believe that more extraverted individuals often interact with large groups of people. Two of the moderately difficult to endorse items were EXT8 *am skilled at dealing with people* and EXT3 *start conversations* suggest that most of the students felt relatively comfortable controlling a conversation with others. The easiest to endorse items, EXT6 *talk a lot* and EXT2 *like to talk to a lot of people*, suggested that the students enjoyed socializing and talking with others.

```
----------------------------------------------------------------------
Persons with more |  More difficult items
extraversion      |
    3             +
                  |
                  |
              XX  |
    2             +
             XXX  |
             XXX S|T
                  |
    1         XX  +  EXT7: Feel at ease in a large group
              XX  |S EXT1: Am the life of the party
                  |  EXT5: Like to be the center of attention
               X  |
               X  |
    0         XX  +M EXT8: Am skilled at dealing with people
              XX M|  EXT3: Start conversations
                  |  EXT4: Like to draw attention to myself
           XXXXX  |S EXT6: Talk a lot
   -1             +
                  |  EXT2: Like to talk to a lot of people
                  |T
                 S|
   -2             +
               X  |
               X  |
              XX  |
   -3             +
                  |
                 T|
               X  |
   -4             +
               X  |
                  |
                  |
   -5             +
Persons with      |
less extraversion |  Easier items
----------------------------------------------------------------------
```

Figure 3. Item-person response map for the Extraversion instrument.
Note: M = mean, S = one standard deviation, T = two standard deviations, X = one student, N = 29.

Table 10 shows the loading of items from the Extraversion measure. Items EXT1, EXT5, and EXT7 had positive loadings, and items EXT2, EXT3, EXT4, EXT6, and EXT8 had negative loadings. Closer examination of the items suggest that they are measuring different aspects of the construct of extraversion, as the items that loaded positively refer to interaction with groups of people, while the items that loaded negatively refer to interaction in general. These results are similar to those found in Leeming (2014), and in a similar fashion to that study, the items were assumed to be a part of a unidimensional construct.

Table 10. *Rasch PCA of Item Residuals for the Extraversion Items*

Item	Loading	Measure	Infit MNSQ	Outfit MNSQ
EXT5	.84	.82	1.47	1.37
EXT1	.77	.71	1.20	1.43
EXT7	.01	1.09	1.21	1.31
EXT4	-.70	-.46	.46	.47
EXT3	-.68	-.30	.56	.58
EXT2	-.48	-1.18	1.14	1.10
EXT8	-.38	.12	.65	.63
EXT6	-.07	-.79	.98	1.00

After the Rasch analysis of the Extraversion measure, descriptive statistics were examined to determine whether the results were normally distributed. The results, using the logits from the Rasch analysis, can be seen in Table 11. Levels of skewness and kurtosis were in acceptable ranges. Person measures were also converted to z-scores and all fell well within the range of -3.29 to 3.29, showing that results for the measure were normally distributed with no outliers.

Table 11. *Descriptive Statistics for the Extraversion Rasch Person Measures*

	Extraversion
M	-1.30
SE	.33
95% CI	[-.8, .54]
SD	1.76
Skewness	-.67
SES	.43
Kurtosis	-.31
SEK	.85

Note. N = 29.

Willingness to Communicate

The eight items on the WTC questionnaire were entered into a Rasch rating scale analysis. The measure used a four-point Likert scale which was marked (1 = *Strongly disagree,* 2 = *Disagree,* 3 = *Agree,* 4 = *Strongly agree*). The Likert-scale category functioning was examined, and can be seen in Table 12. However, the minimum of 10 observations per category was not met, as only one student used *Strongly disagree* on one item. Therefore, the scale was reduced to three points, combining the first two scales into one, labelled *Disagree*. After this reduction, the minimum count was 64 for the category of *disagree*. There was a regular distribution of observations across categories, and the observed averages of the persons increased with the rating scale category. The outfit MNSQ statistic was below 2.00 for all categories. The Andrich thresholds advanced, and there was a relationship between the measures and the average expected ratings. Separation between categories was greater than the required 1.4 logits for a three-point scale, and none were larger than 5.0 (Linacre, 2002). The category characteristic curves can also be seen in Figure 4, which demonstrates how the rating scale accords with the model expectations. After examining these results, it was decided that the Likert-scale was performing in a satisfactory manner.

Table 12. *Category Structure Functioning for Willingness to Communicate*

Category	Count	(%)	Infit MNSQ	Outfit MNSQ	Structure measure	Category measure
1 Disagree	64	17	1.10	1.09	None	(-3.06)
2 Agree	147	51	.89	.96	-1.95	.00
3 Strongly agree	70	27	.97	.95	1.95	(3.06)

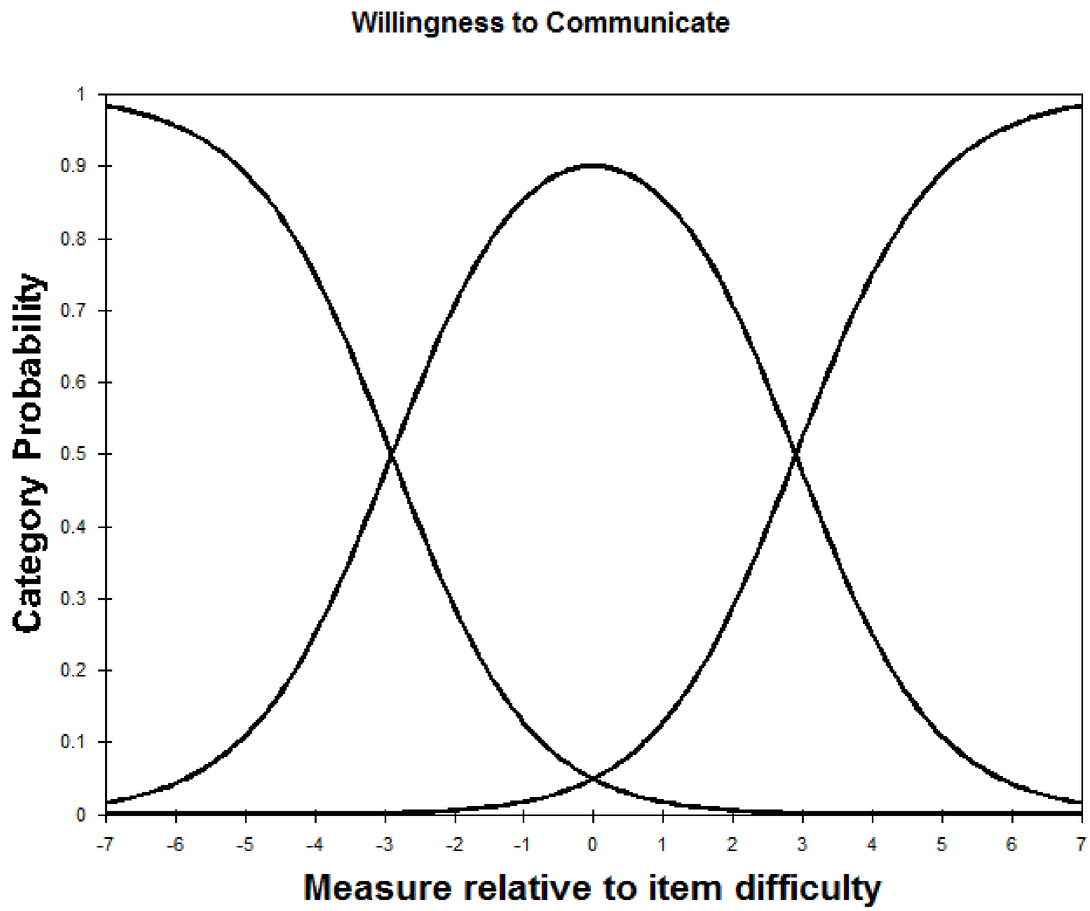

Figure 4. Rating scale performance of WTC measured on a 3-point Likert scale.

Table 13 shows the item fit statistics for the 9 items, which shows that some of the items were not within the acceptable infit MNSQ range of .62 to 1.38 (McNamara, 1996). One reason for this could be due to the small sample size (Bond & Fox, 2015). Items WTC3, *talk about my future plans,* and item WTC5, *disagree with other speakers*, slightly underfit the model, therefore - (infit MNSQ = 1.66 and 1.41, respectively; therefore, person measures with and without these items were exported to SPSS and a bivariate correlation analysis was run. The Pearson correlation was .49 ($p < .001$), therefore it was decided to remove them, as the misfitting responses to these items might have been a result of the content (talking about the future), and a desire to avoid potential conflict (disagreeing with other students) rather than students' levels of WTC. Items WTC4, *give reasons to support my opinions*, and WTC8, *bring up a new topic*, overfit the model (infit MNSQ = .53 and .56, respectively). This overfit was probably the result of the lack of independence among the items, due to the similarity of items, as six of the original nine items involved the language functions that were part of the course objectives. Because overfit has minor implications for measurement (Bond & Fox, 2015), I decided to retain these items for the subsequent analyses.

Table 13. *Rasch Item Statistics for the Willingness to Communicate Items*

Item	Measure	*SE*	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt-measure correlation
WTC1	-.61	.39	.79	-.74	1.27	.90	.64
WTC2	-.61	.39	.70	-1.17	.69	-1.02	.64
WTC3	1.16	.38	1.66	2.45	1.63	2.23	.65
WTC4	-.15	.39	.56	-1.91	.52	-1.95	.65
WTC5	1.02	.38	1.41	1.61	1.36	1.37	.65
WTC6	1.16	.38	1.16	.73	1.14	.62	.65
WTC7	-.30	.39	.98	.01	.93	-.13	.64
WTC8	-1.24	.40	.53	-2.10	.45	-1.77	.61
WTC9	.45	.39	1.06	.30	1.00	.11	.64

The Wright map (Figure 5) shows the students on the logit scale on the far left of the figure. The students are displayed as Xs according to their Rasch WTC measures, with students with higher levels of extraversion toward the top of the map and persons with lower levels of WTC toward the bottom of the map. The items are displayed on the right side of the figure according to their difficulty estimates, with the items that were more difficult to agree with at the top of the map and the items that were easier to agree with at the bottom. The average person measure was 1.00, indicating that the items were relatively easy for students to endorse, which is probably a result of the questionnaire being administered at the conclusion of the course's second semester. This can be seen in the Wright map, which shows that most of the items were easy for the students to endorse.

The items that were the most difficult to endorse were WTC6, *ask questions about other people's experiences* and WTC4, *give reasons to support my opinions.* This could be because such kinds of activities require students to generate content and language at the same time. In contrast, the easiest to endorse items, WTC8, *bring up a new topic*, and WTC9, *suggest a new topic* were based around a single set phrase ("let's talk about {topic}" and "why don't we discuss {topic}") and introduced fairly early in the course.

The willingness to communicate measure had an item reliability of .74 and a person reliability of .76, item separation of 1.68, and person separation of 1.78. While the reliability estimates were fair, the level of separation between both the items and persons were poor (Fisher, 2007). The Rasch PCA of item residuals analysis showed that 50.2% of the variance was explained by the Rasch model, barely surpassing

Linacre's (2015) criterion of 50%. 12.5% of the variance (eigenvalue = 1.76) was explained by the items, and 15.2% of the variance was explained by the first residual contrast, with an eigenvalue of 2.13, meeting Linacre's (2015) criterion of under 3.0.

```
----------------------------------------------------------------------
Persons with     |  More difficult items
more WTC         |
    5        X   +
                 |
                 |
    4      XXX   +
                 |
            XX   |
    3           S+
         XXXX    |
                 |
    2      XXX   +  WTC6: Ask questions about other people's
                 |         experiences.
                 |T
                 |
    1       XX  M+S
                 |
                 |  WTC4: Give reasons to support my opinions.
    0 XXXXXXXX   +M WTC7: Be the first person to speak.
                 |  WTC1: Talk about things I like or I don't like.
                 |  WTC2: Talk about my past experiences.
                 |  WTC9: Suggest a change of topic.
                 |
   -1        X  S+S
                 |  WTC8: Bring up a new topic.
                 |T
   -2      XXX   +
                 |
             X   |
   -3            +
             X  T|
                 |
   -4            +
Persons with     |
less WTC         |  Easier items
----------------------------------------------------------------------
```

Figure 5. Item-person response map for the Willingness to Communicate instrument. *Note:* M = mean, S = one standard deviation, T = two standard deviations, X = one student, *N* = 29.

Table 14 shows the loading of items from the WTC measure. Items WTC5, WTC6, and WTC7 had positive loadings, and items WTC1, WTC2, WTC3, and WTC4 had negative loadings. Closer examination of the items suggest that they are measuring different aspects of the construct of WTC, as the items that loaded positively refer to turn-taking during a group discussion, while the items that loaded negatively refer to the content of a speaking turn. Given that these were behaviors that students could voluntarily engage in during the group discussions held in class, the items were assumed to be a part of a unidimensional construct.

Table 14. *Rasch PCA of Item Residuals for the Willingness to Communicate Items*

Item	Loading	Measure	Infit MNSQ	Outfit MNSQ
WTC8	.83	-1.21	.53	.41
WTC9	.83	-.19	1.15	1.23
WTC7	.04	.00	1.32	1.24
WTC2	-.50	-.39	.77	.61
WTC4	-.49	.20	.86	.69
WTC1	-.46	-.39	.93	1.14
WTC6	-.21	-.21	1.11	.96

After the Rasch analysis of the WTC measure, descriptive statistics were examined to determine whether the results were normally distributed. The results, using the logits from the Rasch analysis, can be seen in Table 15. Levels of skewness and kurtosis were in acceptable ranges. Person measures were also converted to z-scores and all fell well within the range of -3.29 to 3.29, showing that results for the measure were normally distributed with no outliers.

Table 15. *Descriptive Statistics for the Willingness to Communicate Rasch Person Measures*

	Extraversion
M	1.00
SE	-.40
95% CI	[.18, 1.82]
SD	2.17
Skewness	-1.57
SES	.43
Kurtosis	-.74
SEK	.85

Note. *N* = 29.

Perceived Fluency

The expert raters' scores of student recordings were entered into a Rasch FACETS analyses. The criterion for selecting the recordings for use in the expert rating sessions was to select the students with the largest difference in their mean length of run measure. Two students were selected from each class, so that a range of proficiency levels could be included in the study, for a total of eight students. The mean length of run values for the highest and lowest performance, in addition to the difference, can be seen in Table 16.

Table 16. *Mean Length of Run Measures Used in Perceived Fluency Ratings*

Student	Level	Highest MLR	Lowest MLR	Difference
1	1	2.97	1.77	1.20
2	1	4.00	2.33	1.67
3	2	4.26	2.72	1.54
4	2	4.77	3.29	1.48
5	3	4.64	3.20	1.44
6	3	4.39	2.81	1.58
7	4	4.17	2.85	1.32
8	4	4.39	2.50	1.89
M	—	4.20	2.68	1.52

Note. MLR = mean length of run.

Expert ratings of perceived fluency were collected at a rating session after all the data had been collected. These ratings were analyzed with a many-facets Rasch analysis. The Likert-scale category functioning was examined for and is shown in Table 17. While the measure originally adopted a five-point scale, as only one rater used the lowest category, *speech is so halting and fragmentary that conversation is impossible* once, the two lowest descriptors were collapsed into one category, named *disfluent speech* in Table 17. The Rasch-Andrich threshold values showed that the scale functioned satisfactorily as a four-point scale.

Table 17. *Category Structure Functioning for Perceived Fluency*

Category	Count	(%)	Outfit MNSQ	Rasch-Andrich thresholds	*SE*
2. Disfluent speech	28	29	1.0	None	—
3. —	36	38	.9	-7.32	.41
4. —	19	20	.9	-.65	.53
5. Fairly fluent speech	13	14	.8	7.97	.65

Figure 6 shows the map of the Rasch estimates of perceived levels of fluency, and they were centered around a mean of approximately -2.3 logits. The first column shows the logit scale ranging from -10 to +11, and the second column shows the raters' levels of severity, with the strictest raters at the top. The third column shows the students' levels of perceived fluency with the most able students at the top. The fourth column shows the students MLR levels for the two performances, with the slowest performance at the top, as the lower scores on this measure made it more difficult. The final column shows how the four-point Likert scale compares to the logit measures.

Looking at Figure 6 reveals that the raters did vary in their severity. The students were more spread out and ranged from -13 to 7 logits. Their ratings on highest and lowest performance also significantly separated indicating that the raters were able to determine which of the performances were the most fluent. Therefore, despite the relatively small differences in the student mean length of run measures summarized in Table 16, the raters were able to distinguish the stronger performance from the weaker.

```
+-----------------------------------------------+
|Measure|-Raters|+Examinee|-Performance |Scale|
|-------+-------+---------+-------------+-----|
|  11   +       +         +             + (5) |
|       |       | 5       |             |     |
|  10   +       +         +             +     |
|   9   +       +         +             +     |
|       |       |         |             |     |
|   8   +       +         +             + --- |
|   7   +       +         +             +     |
|       |       |         |             |     |
|   6   +       + 2       +             +     |
|   5   +       +         +             +     |
|       |       |         |             |     |
|   4   +       +         +             +     |
|   3   +       +         +             +  4  |
|       |       |         |             |     |
|   2   +       +         + Lowest MLR  +     |
|   1   + 2  4  +         +             +     |
|       | 1  5  |         |             |     |
*   0   *       *         *             *     *
|  -1   +       +         +             + --- |
|       |       |         |             |     |
|  -2   + 3  6  + 1  4    + Highest MLR +     |
|  -3   +       +         +             +     |
|       |       |         |             |     |
|  -4   +       +         +             +  3  |
|  -5   +       + 8       +             +     |
|       |       |         |             |     |
|  -6   +       +         +             +     |
|  -7   +       +         +             +     |
|       |       |         |             | --- |
|  -8   +       + 6       +             +     |
|  -9   +       +         +             +     |
|       |       | 3  7    |             |     |
| -10   +       +         +             + (2) |
|-------+-------+---------+-------------+-----|
|Measure|-Raters|+Examinee|-Performance |Scale|
+-----------------------------------------------+
```

Figure 6. FACET map for perceived fluency. MLR = mean length of run.

Table 18 shows the results from the FACETS analysis for the raters. The raters are ranked according to severity, as shown by the *measure*, with the most severe raters at the top of the table. The fit statistics show that all of the raters fit the Rasch model, although some of the lower values suggest that some raters did not make full use of the scale. The fixed chi-square value was 24.7 with 5 degrees of freedom and significant at $p > .00$, indicating that there were significant differences in the severity of the raters.

Table 18. *Rater Severity for Perceived Fluency*

Rater	Observed count	Measure	Model *SE*	Infit MNSQ	Infit ZSTD
2	16	1.36	.72	.80	-.3
4	16	1.36	.72	.80	-.4
5	16	.87	.69	.71	-.8
1	16	.40	.67	1.17	.6
3	16	-1.99	.72	.84	-.2
6	16	-1.99	.72	1.15	.4

Note. Separation = 1.8; Reliability (reliability in the severity of the raters) = .76; Fixed (all same) chi square = 24.7, $df = 5$, $p = .00$.

Table 19 shows the results from the FACETS analysis for the students whose levels of fluency were rated. The students are ranked in order of most to least fluent. One student overfit the model due to one rater assigning higher scores than the other raters for both performances. The fixed chi-square value was 476.8 with 7 degrees of freedom and significant at $p > .00$, indicating that there were significant differences in the abilities of the students.

Table 19. *Student Levels of Perceived Fluency*

Student	Observed count	Measure	Model *SE*	Infit MNSQ	Infit ZSTD
5	12	10.91	.94	.83	-.30
2	12	6.05	.86	.74	-.30
1	12	-2.29	.76	.50	-1.10
4	12	-2.29	.76	.65	-.60
8	12	-4.51	.73	1.34	.90
6	12	-7.97	.76	1.74	-1.90
3	12	-9.29	.89	.78	-.20
7	12	-9.29	.89	.50	-.90

Note. Separation = 8.26; Reliability (reliability in the differences of students' abilities) = .99; Fixed (all same) chi square = 476.8, *df* = 7, *p* = .00.

Table 20 shows the Rasch fit statistics for the MLR levels in the fluency recordings. The results indicate that judges perceived the performances with slower rates of MLR as less fluent than those with faster rates. The average difference was between the faster and slower rates was approximately 1.5 syllables per run. It should be noted however, that both measures underfit the model due to the low number of items. The fixed chi-square value was 36.9 with 1 degree of freedom and significant at $p > .00$, indicating that there were significant differences in the performances.

Table 20. *Item Difficulty of Highest and Lowest Levels of Perceived Fluency*

MLR	Observed count	Measure	Model *SE*	Infit MNSQ	Infit ZSTD
Lowest	48	1.76	.44	1.00	.0
Highest	48	-1.76	.38	.84	-.6

Note. Separation = 4.18; Reliability (reliability in the perceived differences of the fluency levels) = .95; Fixed (all same) chi square = 36.9, *df* = 1, *p* = .00, MLR = mean length of run.

Summary

In this chapter, I have discussed the measures used in the study in more detail, and outlined how they were examined for reliability and validity. Specifically, I have described how the CAF variables were checked by a second researcher who coded a small portion of the data. The results of these analyses showed that the CAF measures could be quantified to a relative degree of reliability, although it should be noted that background noise created some difficulty in the measurement of the length of pauses. There was consistency, however, in the number of pauses that were over the length of .25 seconds.

Following this, the dimensionality and model fit of the extraversion and WTC measures obtained from the rating-scale Rasch model were described. Finally, the dimensionality, model fit, and inter-rater reliability of the ratings of students' perceived fluency obtained from the many-facets Rasch analysis was also described. All of these measures met the established criteria for use in the study.

CHAPTER 5

RESULTS

In this chapter, I present the results of the quantitative analyses in order to answer the research questions outlined in Chapter 2. First, the processes used for screening the data are explained. Next, I present each research question in turn, and explain the results of each analysis. Following on from this chapter, in Chapter 6 I discuss these results in more detail and attempt to interpret their results in terms of the context of the study and with reference to existing research and foreign language teaching pedagogy.

Missing Cases

As this study was conducted in a one-semester course, most students had one or more absences during the data collection period. One strength of using MLM is that the data can be interpreted as long as each participant has at least one observation, assuming that the data are missing at random (Raudenbush & Bryk, 2002). Therefore, no cases were removed from the model during the main analyses outlined below. For the monologue speaking tests, there were 16 missing observations, composing 12.5% of the total data set. Of the 32 students, four missed two tests, and eight missed one test. For the 3/2/1 speaking activity lessons, there were 29 missing observations, making up 11.4% of the total data. Of the 32 students, one missed three lessons, nine missed two lessons, and eight missed one lesson. The majority of the absences were due to tardiness, illnesses, or extracurricular commitments. Attendance in the class was fairly stable across the semester, as can be seen in Table 21.

Table 21. *Number of Absences During Data Collection*

Data collection lessons	Number of absences	Number of students tested
Lesson 3	2	30
Lesson 4 (Monologue Test 1)	3	29
Lesson 6	5	27
Lesson 7	5	27
Lesson 8 (Monologue Test 2)	7	25
Lesson 10	4	28
Lesson 11	1	31
Lesson 12 (Monologue Test 3)	2	30
Lesson 14 (Monologue Test 4)	4	28

Topics

To investigate whether the topics were of a relatively equal level of difficulty for the students, the confidence intervals for each topic were examined for both the mean length of run, the mean length of pause, the articulation rate, and the phonation/time ratio. As can be seen in Table 22, the confidence intervals for the mean length of run in all of topics were overlapping. While overlapping confidence intervals do not indicate non-significant differences between mean, inspection of this data shows that the intervals are relatively close together, and similarly range from 2.70 to 4.43 syllables per run. Inspection of the means also shows a similar range of values, moving from a low of 3.24 to a high of 3.6 syllables per run.

Table 22. *Mean Length of Run Descriptive Statistics and 95% Confidence Intervals for Topics*

Topic	*N*	*M*	*SE*	*SD*	95% CI
1	10	3.24	.24	.76	[2.70, 3.79]
2	13	3.76	.15	.53	[3.43, 4.07]
3	10	3.36	.13	.40	[3.08, 3.65]
4	11	3.57	.25	.83	[3.01, 4.12]
5	8	3.65	.29	.82	[2.96, 4.34]
6	10	3.25	.21	.67	[2.77, 3.73]
7	10	3.72	.32	1.0	[3.00, 4.43]
8	10	3.49	.26	.81	[2.90, 4.07]
9	15	3.34	.18	.71	[2.95, 3.74]
10	15	3.38	.22	.85	[2.91, 3.85]

Note. CI = Confidence interval.

Similarly, Table 23 shows that the confidence intervals for the mean length of pause in all of the topics were overlapping, suggesting that they were of equal levels of difficulty for the students. Inspection of this data shows that the intervals are relatively close together, and similarly range from .91 to 1.86 seconds. Inspection of the means also shows a similar range of values, moving from a low of 1.15 to a high of 1.44 seconds.

Table 23. *Mean Length of Pause Descriptive Statistics and 95% Confidence Intervals for Topics*

Topic	*N*	*M*	*SE*	*SD*	95% *CI*
1	10	1.31	.14	.43	[.99, 1.62]
2	13	1.17	.08	.28	[1.00, 1.33]
3	10	1.31	.13	.43	[1.01, 1.62]
4	11	1.28	.11	.37	[1.03, 1.52]
5	8	1.44	.18	.50	[1.03, 1.86]
6	10	1.32	.12	.38	[1.05, 1.60]
7	10	1.17	.05	.17	[1.06, 1.29]
8	10	1.17	.12	.37	[.91, 1.43]
9	15	1.33	.10	.37	[1.12, 1.53]
10	15	1.15	.11	.42	[.92, 1.39]

Note. CI = Confidence interval.

Likewise, Table 24 shows that the confidence intervals for the articulation rate in all of topics were overlapping, suggesting that they were of equal levels of difficulty for the students. Inspection of this data shows that the intervals are relatively close together, and similarly, they range from 132.77 to 219.29 syllables. Inspection of the means also shows a similar range of values, moving from a low of 164.79 to a high of 195.14 syllables.

Table 24. *Articulation Rate Descriptive Statistics and 95% Confidence Intervals for Topics*

Topic	*N*	*M*	*SE*	*SD*	95% *CI*
1	10	172.21	8.60	27.21	[152.75, 191.67]
2	13	171.17	6.36	22.93	[157.31, 185.03]
3	10	170.11	5.91	18.69	[156.74, 183.48]
4	11	183.40	10.25	34.01	[160.55, 206.24]
5	8	181.74	15.88	44.91	[144.19, 219.29]
6	10	164.79	14.15	44.75	[132.77, 196.80]
7	10	195.14	10.58	33.47	[171.20, 219.08]
8	10	184.83	7.18	22.69	[168.60, 201.06]
9	15	186.56	13.92	53.90	[156.71, 216.40]
10	15	184.63	7.02	27.21	[169.57, 199.70]

Note. CI = Confidence interval.

Finally, Table 25 shows that the confidence intervals for the phonation/time ratio in all of topics were overlapping, suggesting that they were of equal levels of difficulty for the students. Inspection of this data shows that the intervals are relatively close together, and similarly range from .38 to .58 seconds. Inspection of the means also shows a similar range of values, from a low of .46 to a high of .54 seconds.

Table 25. *Phonation/Time Ratio Descriptive Statistics and 95% Confidence Intervals for Topics*

Topic	*N*	*M*	*SE*	*SD*	95% *CI*
1	10	.47	.04	.12	[.38, .55]
2	13	.54	.02	.06	[.51, .58]
3	10	.49	.03	.08	[.43, .55]
4	11	.49	.02	.07	[.44, .54]
5	8	.47	.03	.08	[.40, .54]
6	10	.50	.03	.09	[.43, .56]
7	10	.49	.02	.06	[.45, .54]
8	10	.49	.04	.12	[.41, .58]
9	15	.46	.03	.11	[.40, .53]
10	15	.50	.03	.12	[.44, .57]

Note. CI = Confidence interval.

Long-Term Fluency Measures

Before running the MLM analyses, descriptive statistics were calculated for the fluency variables from the monologue tests used in the study. As measures of long-term fluency, the mean length of run and mean length of pause were examined, as development in one measure in combination with stability or growth in the other has been argued to show fluency development (De Jong & Perfetti, 2011). A total of five fluency measures were investigated: the mean length of run, the mean length of pause, the articulation rate, the phonation/time ratio, and the mid-clause pause ratio. For the mean length of run (see Table 26), skewness and kurtosis were within acceptable limits for all of the tests, indicating that the data were normally distributed.

Table 26. *Descriptive Statistics for Mean Length of Run in Monologue Tests*

	N	*M*	*SE*	*SD*	Skewness	*SES*	Kurtosis	*SEK*
Test 1	29	3.47	.13	.74	.59	.42	-.17	.83
Test 2	25	3.46	.14	.71	.52	.46	.25	.87
Test 3	30	3.51	.16	.86	-.19	.43	-.62	.85
Test 4	28	3.44	.14	.74	.86	.44	.36	.86

For the mean length of pause measures (see Table 27), skewness and kurtosis were within acceptable limits for all of the tests. This indicated that the data were normally distributed.

Table 27. *Descriptive Statistics for Length of Pause in Monologue Tests*

	N	*M*	*SE*	*SD*	*Skewness*	*SES*	Kurtosis	*SEK*
Test 1	29	1.35	.08	.42	.82	.43	-.26	.83
Test 2	25	1.26	.07	.37	.90	.46	.46	.89
Test 3	30	1.26	.06	.33	.86	.43	.33	.85
Test 4	28	1.16	.07	.37	.78	.45	.95	.87

For the articulation rate (see Table 28), skewness and kurtosis were within acceptable limits for all of the tests, indicating that the data were normally distributed.

Table 28. *Descriptive Statistics for Articulation Rate in Monologue Tests*

	N	*M*	*SE*	*SD*	*Skewness*	*SES*	Kurtosis	*SEK*
Test 1	29	174.86	5.06	22.65	.33	.51	-4.43	.99
Test 2	25	173.49	5.79	25.91	-.02	.51	-.81	.99
Test 3	30	175.17	5.23	23.38	-.37	.51	-1.03	.99
Test 4	28	168.98	5.89	26.35	.65	.51	.78	.99

For the phonation/time ratio (see Table 29), skewness and kurtosis were within acceptable limits for all of the tests, indicating that the data were normally distributed.

Table 29. *Descriptive Statistics for Phonation/Time Ratio in Monologue Tests*

	N	*M*	*SE*	*SD*	*Skewness*	*SES*	Kurtosis	*SEK*
Test 1	29	.48	.02	.08	-.19	.43	.06	.83
Test 2	25	.50	.02	.09	-.41	.46	.99	.89
Test 3	30	.49	.02	.09	-.22	.43	-.67	.85
Test 4	28	.52	.02	.10	-.37	.45	.71	.87

The measurements of the mid-clause pause ratio (see Table 30), skewness and kurtosis were within acceptable limits for all of the tests, indicating that the data were normally distributed.

Table 30. *Descriptive Statistics for Mid-Clause Pause Ratio*

	N	*M*	*SE*	*SD*	*Skewness*	*SES*	Kurtosis	*SEK*
Test 1	29	.66	.02	.11	.16	.43	-.58	.85
Test 2	25	.62	.03	.13	.84	.46	.68	.90
Test 3	30	.69	.02	.10	-.31	.43	-.10	.84
Test 4	28	.69	.02	.12	-.77	.44	1.22	.86

Complexity Measures

This study included two measures of syntactic complexity, the ratio of clauses to the total number of AS-units, and the average number of words per AS-unit. For lexical complexity, the MTLD was used. As can be seen in Table 31 for the measures of syntactic complexity (the number of clauses per AS-unit), skewness and kurtosis were within acceptable limits for all of the tests, indicating that the measure was normally distributed.

Table 31. *Descriptive Statistics for Clauses per AS-Unit in Monologue Tests*

	N	*M*	*SE*	*SD*	*Skewness*	*SES*	Kurtosis	*SEK*
Test 1	29	1.73	.12	.53	.37	.51	-1.58	.99
Test 2	25	1.67	.08	.36	.85	.51	-3.10	.99
Test 3	30	1.83	.08	.36	.43	.51	.06	.99
Test 4	28	1.74	.09	.39	.73	.51	.34	.99

In contrast, the number of words per AS-unit, as seen in Table 32, was both skewed and leptokurtic. For this reason, the number of clauses per AS-unit was chosen as the independent variable for the MLM analysis.

Table 32. *Descriptive Statistics for Words per AS-Unit in Monologue Tests*

	N	*M*	*SE*	*SD*	*Skewness*	*SES*	Kurtosis	*SEK*
Test 1	29	8.98	.52	2.83	1.45	.43	2.64	.85
Test 2	25	9.09	.58	2.92	1.69	.46	5.22	.90
Test 3	30	9.29	.50	2.74	1.34	.43	2.34	.83
Test 4	28	10.49	.86	4.57	2.58	.44	9.43	.86

The statistics for lexical complexity, as measured by the MTLD, are shown in Table 33. While three of the tests met the assumptions of normality, the results for the second test did not. In this test, one student from class 3 outperformed the other students by more than 20 points on the MTLD scale, due to the listing of a series of nouns (foods her mother cooks) in her speech. This score was reduced to the mean plus two standard deviations (Field, 2009, p. 153). This changed the mean from 33.58 to 33.11 and the standard deviation from 9.19 to 7.83.

Table 33. *Descriptive Statistics for the Measure of Textual Lexical Diversity in Monologue Tests*

	N	*M*	*SE*	*SD*	*Skewness*	*SES*	Kurtosis	*SEK*
Test 1	29	37.75	2.34	10.50	.89	.51	-1.21	.99
Test 2	25	33.11	1.75	7.83	.82	.51	.45	.99
Test 3	30	36.29	1.78	7.95	-.05	.51	.56	.99
Test 4	28	39.02	2.29	10.28	.54	.51	-.63	.99

Finally, although the MTLD is less influenced by text length as compared to other measures such as the type-token ratio (Jarvis, 2013), correlations were run between text length that was measured by total number of words and MTLD values to determine whether the shorter speeches in the study had inflated MTLD values. The results, presented in Table 34, show that although there appears to be a pattern of weak positive relationships between the two variables, none were statistically significant. It should be mentioned that this lack of significance could be due to the small sample size of the study. However, for the purposes of this study, MTLD is accepted as a measure of lexical diversity.

Table 34. *Correlations Between Speech Length and the Measure of Textual Lexical Diversity*

	N	*r*	*p*
Test 1	29	.13	.51
Test 2	25	.22	.29
Test 3	30	.20	.30
Test 4	28	.23	.24

Accuracy Measures

The following measures of accuracy were used in this study: the total number of errors, the number of words per error, and the ratio of error-free clauses to the total number of clauses. As can be seen in Table 35, the results for the total number of errors were both skewed and leptokurtic, the result of a small number of students giving largely error-free performances combined with the majority of students who made a large number of errors.

Table 35. *Descriptive Statistics for Total Number of Errors in Monologue Tests*

	N	*M*	*SE*	*SD*	*Skewness*	*SES*	Kurtosis	*SEK*
Test 1	29	7.79	.87	4.68	.57	.43	-.76	.85
Test 2	25	8.52	.96	4.81	1.36	.46	3.24	.90
Test 3	30	8.07	.70	3.82	-.10	.43	-1.00	.83
Test 4	28	7.39	.62	3.29	.48	.44	.23	.86

In a similar fashion, the results for the number of words per error were also skewed and leptokurtic, as can be seen in Table 36.

Table 36. *Descriptive Statistics for Words per Error in Monologue Tests*

	N	*M*	*SE*	*SD*	*Skewness*	*SES*	Kurtosis	*SEK*
Test 1	29	15.64	2.66	14.33	2.68	.43	7.10	.85
Test 2	25	14.22	2.51	12.56	2.82	.46	8.84	.90
Test 3	30	14.31	2.52	13.83	3.83	.43	17.25	.83
Test 4	28	14.06	1.92	10.15	3.66	.44	16.19	.86

However, an examination of the descriptive statistics for the error-free clause ratio distribution (see Table 37) met assumptions of normality, as the errors made by the majority of students were often concentrated in the same clauses, which led to a wider range of variation. Based on these results, the error-free clause ratio was chosen as the most appropriate accuracy variable for further analysis, as levels of skewness and kurtosis were within acceptable limits for all of the tests, indicating that the data were normally distributed.

Table 37. *Descriptive Statistics for the Error-Free Clause Ratio in Monologue Tests*

	N	*M*	*SE*	*SD*	*Skewness*	*SES*	Kurtosis	*SEK*
Test 1	29	.64	.03	.17	.08	.43	-1.10	.85
Test 2	25	.61	.03	.16	-.10	.46	-.64	.90
Test 3	30	.62	.02	.13	.16	.43	-.44	.83
Test 4	28	.65	.03	.14	-.30	.44	-.62	.86

In-Class Performance Measures

Three measures of student performance in class activities were used in this study: the average number of words repeated in each lesson, the average number of words spoken in each lesson, and the average number of formulaic phrases used in each lesson, counted by type, not token. To qualify for inclusion in the average number of repeated words, the word in question had to be used in all five of the speaking activities. As can be seen in Table 38, skewness and kurtosis were within acceptable limits for all three of the measurements, indicating that they met the assumptions of normality.

Table 38. *Descriptive Statistics for In-Class Performance Measures*

	M	*SE*	*SD*	*Skewness*	*SES*	Kurtosis	*SEK*
Repeated	16.26	.76	4.28	.61	.41	.90	.81
Spoken	865.64	40.22	227.49	.70	.41	1.60	.81
Form	3.06	.11	.61	-.19	.41	-.07	.81

Note. Repeated = average number of repeated words per class; Spoken = average number of words spoken per class; Form = average number of formulaic phrases used per class, *N* = 32.

Assumptions of Multi-Level Modeling

Prior to running the MLM analyses, the data were also screened for multivariate outliers by examining the Mahalanobis distance for each monologue test. As there were six independent variables (the three in-class performance measures, WTC, extraversion, and L1 fluency), the χ^2 at the criterion of $p < .001$ was 22.458. There were no cases above these values, signaling that the data contained no multivariate outliers (Tabachnick & Fidell, 2007, p. 99).

Short-Term Fluency Measures

To measure short-term fluency, student performances on the eight 3/2/1 speaking tasks were examined. Similar to the monologue tests, the skewness and kurtosis levels of mean length of run and mean length of pause were within acceptable limits for most of the 3/2/1 speeches, however, there were several tests with positively skewed results. The outlying scores were therefore reduced to the mean plus two standard deviations, recommended by Field (2009, p. 153). There were two occasions in the 3-minute speeches when the mean length of run was not normally distributed.

This first occurred in Lesson 6, when two students repeated use of phrases from the text (e.g. *kissing in public* and *eating on the train* in the lesson on public rules) resulted in outlying mean length of run scores. After adjustment, the mean was reduced from 3.42 to 3.39 and the standard deviation from .77 to .69. Second, in lesson 7, one student used coordinating conjunctions within runs several times, which also led to an outlying mean length of run on the 2 min speech. After adjustment, the mean was reduced from 3.79 to 3.71 and the standard deviation from .85 to .74. The final descriptive statistics are shown in Table 39. As can be seen, the skewness and kurtosis were within acceptable limits for all of the tests, indicating that the final data were normally distributed.

Table 39. *Descriptive Statistics for Mean Length of Run in 3-Minute Speeches*

	N	*M*	*SE*	*SD*	*Skewness*	*SES*	Kurtosis	*SEK*
Time 1	30	3.63	.14	.77	.61	.43	-.38	.83
Time 2	29	3.77	.14	.76	.16	.43	.74	.85
Time 3	27	3.52	.13	.67	.19	.45	-.27	.87
Time 4	27	3.66	.14	.73	.52	.45	-.42	.87
Time 5	26	4.14	.14	.73	.85	.46	.75	.89
Time 6	28	3.39	.13	.69	.63	.44	.12	.86
Time 7	30	3.37	.13	.71	.48	.43	-.92	.83
Time 8	30	3.31	.11	.60	.34	.43	.12	.83

The descriptive statistics for the mean length of run in the 2-minute speeches in Table 40 show skewness and kurtosis were within acceptable limits, indicating that the data were normally distributed.

Table 40. *Descriptive Statistics for Mean Length of Run in 2-Minute Speeches*

	N	*M*	*SE*	*SD*	*Skewness*	*SES*	Kurtosis	*SEK*
Time 1	30	3.95	.15	.87	.58	.43	.47	.83
Time 2	29	4.13	.16	.87	.29	.43	.45	.85
Time 3	27	3.85	.14	.70	.21	.45	-.59	.87
Time 4	27	4.16	.15	.79	.23	.45	.26	.87
Time 5	26	4.62	.19	.98	.67	.46	.07	.89
Time 6	28	3.88	.21	1.09	.75	.44	-.11	.86
Time 7	30	3.71	.14	.74	.80	.43	-.04	.83
Time 8	30	3.69	.15	.79	.08	.43	.13	.83

Similarly, the descriptive statistics for the mean length of run in the 1-minute speeches in Table 41 show skewness and kurtosis were within acceptable limits, indicating that the data were normally distributed.

Table 41. *Descriptive Statistics for Mean Length of Run in 1-Minute Speeches*

	N	*M*	*SE*	*SD*	*Skewness*	*SES*	Kurtosis	*SEK*
Time 1	30	4.98	.22	1.22	-.32	.43	-1.01	.83
Time 2	29	4.52	.21	1.14	.31	.43	-.58	.85
Time 3	27	4.51	.21	1.08	.85	.45	1.71	.87
Time 4	27	5.03	.23	1.17	.04	.45	-.12	.87
Time 5	26	5.28	.21	1.09	.90	.46	-.02	.89
Time 6	28	4.63	.29	1.35	.87	.44	.08	.86
Time 7	30	4.59	.22	1.23	.62	.43	.15	.83
Time 8	30	4.39	.19	1.04	-.06	.43	-.92	.83

In terms of the mean length of pause, there were several outlying students. There were two reasons for this. First, some students appeared to lose their train of thought which resulted in very long pauses of around five seconds. Second, though less common, occasionally speakers were distracted by their partner or another student. This was also reported in De Jong and Perfetti's fluency, where all long pauses were reduced to the mean plus 2 standard deviations (2011). However, due to the fact that some pauses could be related to language processing issues, only the pauses that were not in normally distributed data sets were reduced. For the three-minute speeches, the data were not normally distributed in lessons 2, 4, and 6. In Lesson 2, the means and standard deviations were reduced from 1.21 and .34 to 1.19 and .28, respectively. In Lesson 4, the mean and standard deviation was reduced from 1.31 and .33 to 1.30 and .29, respectively. In Lesson 6, the mean and standard deviation was reduced from 1.21 and .30 to 1.19 and 0.5, respectively. In total, three students' scores were adjusted. One of these students was from the highest class and two were from the lowest class. The final descriptive statistics for the man pause rate can be seen in Table 42, and show skewness and kurtosis were within acceptable limits, indicating that the data were normally distributed.

Table 42. *Descriptive Statistics for Mean Length of Pause in 3-Minute Speeches*

	N	*M*	*SE*	*SD*	*Skewness*	*SES*	Kurtosis	*SEK*
Time 1	30	1.03	.05	.36	.54	.43	-.49	.83
Time 2	29	1.19	.05	.28	.74	.43	1.27	.85
Time 3	27	1.41	.07	.35	.51	.45	1.02	.87
Time 4	27	1.30	.05	.29	.83	.45	.87	.87
Time 5	26	1.23	.07	.34	.46	.46	-.47	.89
Time 6	28	1.19	.05	.25	.23	.44	-.09	.86
Time 7	30	1.29	.06	.34	.34	.43	-.27	.83
Time 8	30	1.24	.05	.26	-.16	.43	-.13	.83

For two-minute speeches, the score was adjusted for one student from the second highest class, reducing the mean and standard deviation from 1.08 and .29 to 1.07 and .24, respectively. The results can be seen in Table 43, which shows that skewness and kurtosis were within acceptable limits, indicating that the data were normally distributed.

Table 43. *Descriptive Statistics for Length of Pause in 2-Minute Speeches*

	N	*M*	*SE*	*SD*	*Skewness*	*SES*	Kurtosis	*SEK*
Time 1	30	1.10	.05	.26	.82	.43	.19	.83
Time 2	29	.99	.04	.21	.32	.43	-.43	.85
Time 3	27	1.14	.05	.24	.38	.45	-.58	.87
Time 4	27	1.12	.05	.27	.37	.45	-1.02	.87
Time 5	26	1.06	.05	.27	.88	.46	1.40	.89
Time 6	28	1.07	.05	.24	.78	.44	.45	.86
Time 7	30	1.09	.04	.24	.42	.43	.38	.83
Time 8	30	1.04	.03	.17	.16	.43	-.73	.83

The descriptive statistics for the mean length of pause in the 1-minute speeches met the assumptions of normality, which can be seen in Table 44 which shows that show skewness and kurtosis were within acceptable limits, indicating that the data were normally distributed.

Table 44. *Descriptive Statistics for Length of Pause in 1-Minute Speeches*

	N	*M*	*SE*	*SD*	*Skewness*	*SES*	Kurtosis	*SEK*
Time 1	30	.93	.04	.21	.55	.43	-.38	.83
Time 2	29	.85	.03	.17	.16	.43	-.22	.85
Time 3	27	.97	.04	.20	.50	.45	-.60	.87
Time 4	27	.93	.04	.23	.77	.45	-.38	.87
Time 5	26	.84	.04	.19	.53	.46	.06	.89
Time 6	28	.89	.04	.21	.49	.44	.24	.86
Time 7	30	.95	.04	.24	.51	.43	.46	.83
Time 8	30	.88	.03	.18	.53	.43	-.52	.83

Research Question 1: Long-Term Fluency

Research question 1 investigated the degree to which students' long-term fluency developed over the semester. This question was answered by using multi-level linear modeling to determine the growth trajectory of the fluency measures in the monologue speaking tests over four time periods. Because the construct of fluency has more than one component, four separate analyses were run, as improvement in one area of fluency could occur with a decrease in another. In the first model, the mean length of run was the dependent variable, and time was the independent variable. In the second model, the mean length of pause was the independent variable, and time was the independent variable. In the third model, the articulation rate was the independent variable, and time was the independent variable. Finally, in the fourth model, the phonation/time ratio was the independent variable, and time was the independent variable.

To determine whether the fluency measures significantly changed over the course of the semester, two-level HLM models were run. The first analysis conducted for each measure was to inspect the unconditional model, which investigated the development of each measure without any explanatory variables. If the unconditional model shows that there is significant development among the individuals in the study, further analysis is justified. The equation for the unconditional models is:

$$Y_{ti} = \beta_{0ti} + \beta_{1i}(TIME) + r_{ti}$$

In this equation, Y_{ti} is the fluency measure at time t for learner i, β_{0ti} is the initial measure in fluency in the first test, β_{1ti} is the growth rate for learners i, and r_{ti} is the residual associated with learners i at time t. The level 2 unconditional model is shown below. In this equation, the initial measure of fluency and the growth rates are random coefficients. The initial measure of fluency is represented by

$$\beta_{0i} = \gamma_{00} + u_0$$

$$\beta_1 = \gamma_{10} + u_1$$

The unconditional model for mean length of run was run with no predictors and the results are shown in Table 45.

Table 45. *Statistical Results of the Unconditional Model of the Mean Length of Run*

	Fixed effects			
	Coefficient	SE	*t*-ratio	*p*
Intercept (mean length of run growth over time) γ_{10}	-.005	.04	-.12	.91
	Random effects			
	Variance	*df*	$\chi 2$	*p*
Between-test variability (Intercept)	.33	31	172.89	<.0001
Within-test variability	.24			
Reliability (Intercept)		.82		
Intraclass Correlation		.57		

Note. *N* = 32.

The students' mean length of run measures slightly decreased over the four periods of time at -.005 syllables per test, as can be seen in the coefficient for the fixed effects in Table 45. However, this amount of variation was not statistically significant. The random effects do show that there was significant variance in student performance in each test (var (u_1) = .33, $\chi^2(31) = 172.89$, $p < .0001$. The intraclass correlation showed that 57% of the variance in mean length of run in test scores was attributable to when the test was administered, while 43% is attributable to variation in students' performance. In summary, while there was significant group variation in students' performance in the tests, there was no significant growth in the mean length of run measure. These results can also be seen in Figure 7. Inspection of this graph shows that there was a great deal of fluctuation in this measure for most of the students.

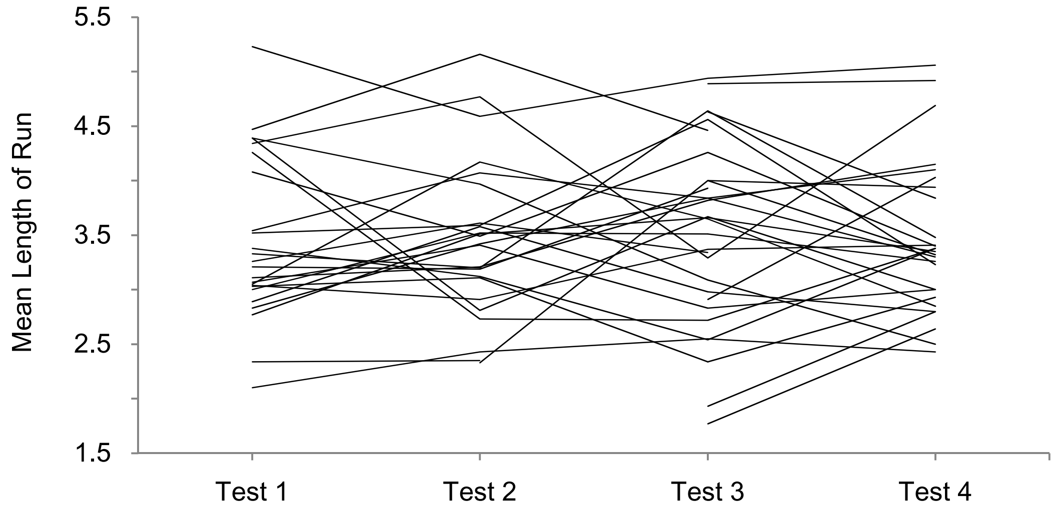

Figure 7. Changes in the mean length of run in monologue speaking tests.

The unconditional model for the mean length of pause was run with no predictors and the results are shown in Table 46.

Table 46. *Statistical Results of the Unconditional Model of the Mean Length of Pause*

	Fixed effects			
	Coefficient	SE	*t*-ratio	*p*
Intercept (mean length of pause growth over time) γ_{10}	-.06	.03	-2.32	.02
	Random effects			
	Variance	*df*	$\chi 2$	*p*
Between-test variability (Intercept)	.06	31	130.64	<.0001
Within-test variability	.07			
Reliability (Intercept)		.75		
Intraclass Correlation		.47		

Note. *N* = 32.

On average, the students' mean length of pause measures slightly decreased over the four periods of time at -.06 seconds per test, as can be seen in the coefficient for the fixed effects in Table 46, and this improvement was significant (p = .02). The results of the random effects also show that there was significant residual variance in student performance in each test (var (u_1) = .06, $\chi^2(31) = 130.64$, $p < .0001$. The intraclass correlation showed that 47% of the variance in mean length of pause in test scores was attributable to when the test was administered, while 53% was attributable to the variance in the students' performance. In summary, there was significant group variation in students' performance on the tests, and there was a small but significant reduction in the mean length of pause measure. These results can also be seen in Figure 8. Inspection of this graph shows that while there is some fluctuation in this variable for some of the students, a cluster of lines at the bottom of the graph move in a downward trend, signaling some degree of reduction in the mean length of pause.

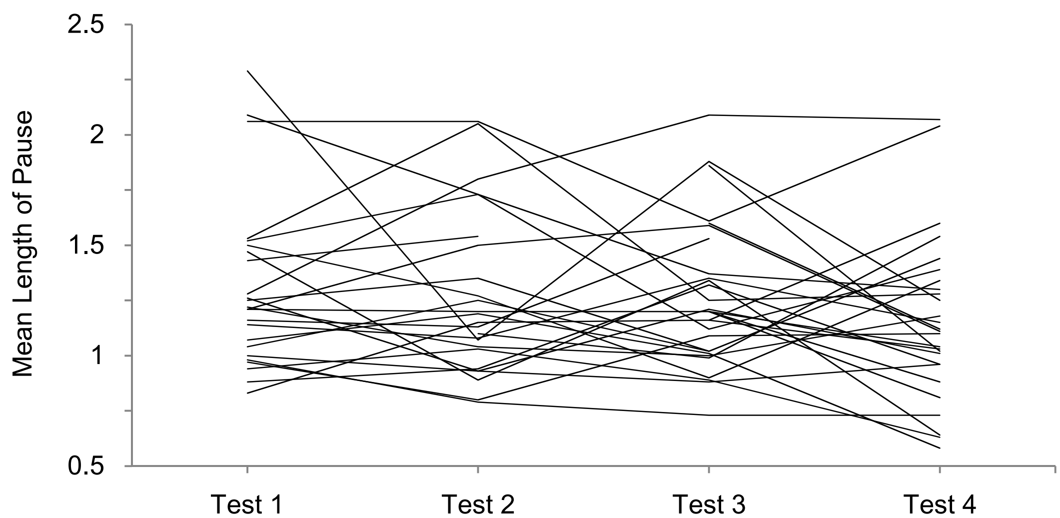

Figure 8. Changes in the mean length of pause in monologue speaking tests.

To investigate the unexplained variance found in the model, each of the independent variables were tested to determine if they could explain any of this group variance. Given the exploratory nature of this study and the small sample size, each independent variable was run in a separate model, the results of which can be seen in Table 47. An example of the equation for these analyses, in this case, student levels of extraversion, is:

$$\beta_{1i} = \gamma_{10} + \gamma_{10}\ (\text{EXTRAVERSION}) + u_{1i}$$

Table 47. *Estimation of Significant Effects of Predictor Variables on Mean Length of Pause*

Fixed effect	Coefficient	*SE*	*t*-ratio	*df*	*p*
WTC	-.00010	.0002	-.37	78	.71
Extraversion	-.00014	.0002	-.49	78	.63
L1 Fluency	-.00031	.0002	-1.41	78	.16
Repeated	.00360	.0042	.86	78	.40
Spoken	.00008	.0001	.90	78	.37
Formulaic	.00243	.0349	.70	78	.49

Note. WTC = Willingness to communicate; Repeated = average number of repeated words per class; Spoken = average number of words spoken per class; Formulaic = average number of formulaic phrases used per class, *N* = 32.

None of the predictor variables had a significant effect in the variation of the significant growth in the mean length of pause, possibly due to the small size of the decrease in the mean length of pause and/or the small n-size of the study.

The unconditional model for the articulation rate was run with no predictors and the results are shown in Table 48.

Table 48. *Statistical Results of the Unconditional Model of the Articulation Rate*

	Fixed effects			
	Coefficient	SE	*t*-ratio	*p*
Intercept (mean length of run growth over time) γ_{10}	-.85	1.83	-.47	.64
	Random effects			
	Variance	*df*	$\chi 2$	*p*
Between-test variability (Intercept)	898.02	31	254.97	<.0001
Within-test variability	401.82			
Reliability (Intercept)		.88		
Intraclass Correlation		.69		

Note. N = 32.

On average, the students' articulation rate slightly decreased over the four periods of time at -.85 syllables per test, as can be seen in the coefficient for the fixed effects in Table 48. However, this amount of variation was not statistically significant. The random effects do show that there was significant variance in student performance in each test (var (u_1) = 898.02, $\chi^2(31) = 254.97$, $p < .0001$. The intraclass correlation showed that 69% of the variance in mean length of run in test scores was attributable to when the test was administered, while 31% was attributable to the variance in the students' performance. In summary, while there was significant group variation in students' performance in the tests, there was no significant growth in the articulation rate. These results can also be seen in Figure 9. Inspection of this graph shows that there was a fair amount of fluctuation in this measure for most of the students.

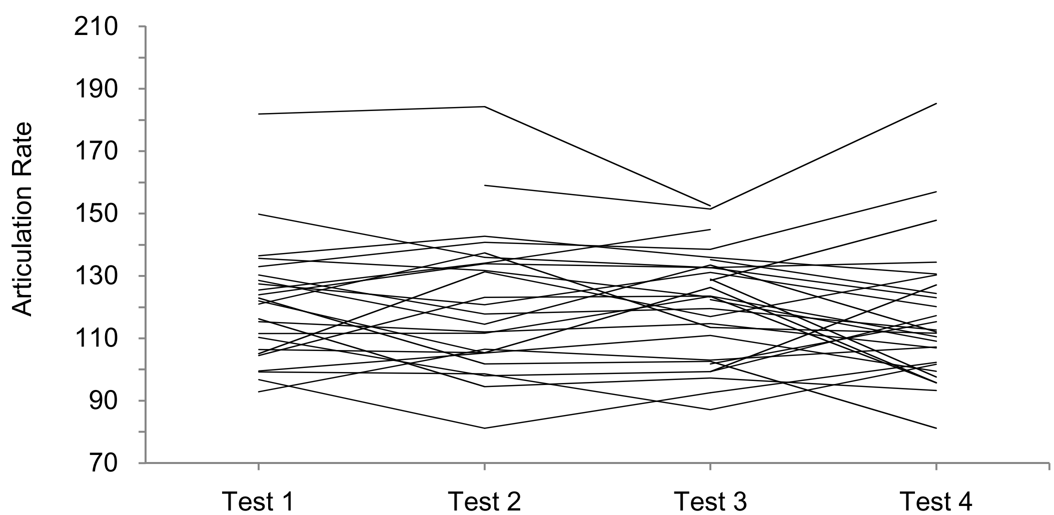

Figure 9. Changes in the articulation rate in monologue speaking tests.

The unconditional model for the phonation/time ratio was run with no predictors and the results are shown in Table 49.

Table 49. *Statistical Results of the Unconditional Model of the Phonation/Time Ratio*

	Fixed effects			
	Coefficient	SE	*t*-ratio	*p*
Intercept (mean length of pause growth over time) γ_{10}	.01	.005	2.09	.04
	Random effects			
	Variance	*df*	$\chi 2$	*p*
Between-test variability (Intercept)	.005	31	175.20	<.0001
Within-test variability	.003			
Reliability (Intercept)		.82		
Intraclass Correlation		.57		

On average, the students' phonation/time ratio slightly increased over the four periods of time by .01 per test, as can be seen in the coefficient for the fixed effects in Table 40, and this improvement was significant (p = .04). The results of the random effects also show that there was significant residual variance in student performance in each test (var (u_1) = .07, $\chi^2(31) = 175.20$, $p < .0001$. The intraclass correlation showed that 57% of the variance in phonation/time ratio in test scores was attributable to when the test was administered, while 43% was attributable to the variance in the students' performance. In summary, there was significant group variation in students' performance on the tests, and there was a small but significant reduction in the mean length of pause measure. These results can also be seen in Figure 10. Inspection of this graph shows that while there is some fluctuation in this variable for some of the students, many of the lines are moving in an upward direction, showing that students were able to spend more time speaking.

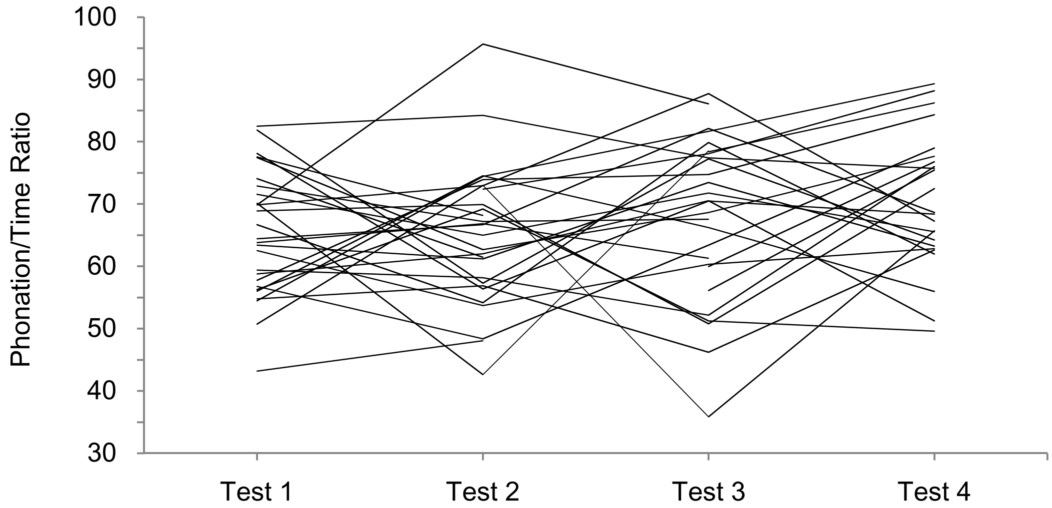

Figure 10. Changes in the phonation/time ratio in monologue speaking tests.

To investigate the unexplained variance found in the model, each of the independent variables were tested to determine if they could explain any of this group variance. Given the exploratory nature of this study and the small sample size, each independent variable was run in a separate model, the results of which can be seen in Table 50. An example of the equation for these analyses, in this case, student levels of extraversion, is:

$$\beta_{1i} = \gamma_{10} + \gamma_{10}\,(\text{EXTRAVERSION}) + u_{1i}$$

Table 50. *Estimation of Significant Effects of Predictor Variables on Phonation/Time Ratio*

Fixed effect	Coefficient	*SE*	*t*-ratio	*df*	*p*
WTC	-.000010	.00006	-.20	78	.84
Extraversion	-.000036	.00007	-.53	78	.60
L1 Fluency	.000057	.00005	-1.41	78	.16
Repeated	.000730	.00101	1.19	78	.24
Spoken	-.000003	.00002	.90	78	.92
Formulaic	-.010791	.00063	-1.71	78	.09

Note. WTC = willingness to communicate; Repeated = average number of repeated words per class; Spoken = average number of words spoken per class; Formulaic = average number of formulaic phrases used per class, *N* = 32.

None of the predictor variables had any significant effect in the variation of the significant growth in the phonation/time ratio. This could have been due to the small size of the growth and/or the small *N*-size of the study.

The unconditional model for the mid-clause pause ration was run with no predictors and the results are shown in Table 51.

Table 51. *Statistical Results of the Unconditional Model of the Mid-Clause Pause Ratio*

	Fixed effects			
	Coefficient	SE	*t*-ratio	*p*
Intercept (mean length of run growth over time) γ_{10}	.01	.007	1.83	.07
	Random effects			
	Variance	*df*	$\chi 2$	*p*
Between-test variability (Intercept)	.004	31	77.51	<.0001
Within-test variability	.009			
Reliability (Intercept)		.59		
Intraclass Correlation		.40		

Note. N = 32.

On average, the students' mid-clause pause ratios slightly improved over the four periods of time by .01 per test, as can be seen in the coefficient for the fixed effects in Table 51. However, this amount of variation was not statistically significant. The random effects do show that there was significant variance in student performance in each test (var (u_I) = .004 $\chi^2(31) = 77.51$, $p < .0001$. The intraclass correlation showed that 40% of the variance in the mid-clause pause ratio on test scores was attributable to when the test was administered, while 60% was attributable to the variance in the students' performance. In summary, while there was significant group variation in the students' performances, there was no significant growth in the mid-clause pause ratio. These results can also be seen in Figure 11, which shows that there was a good deal of fluctuation in this measure for most of the students.

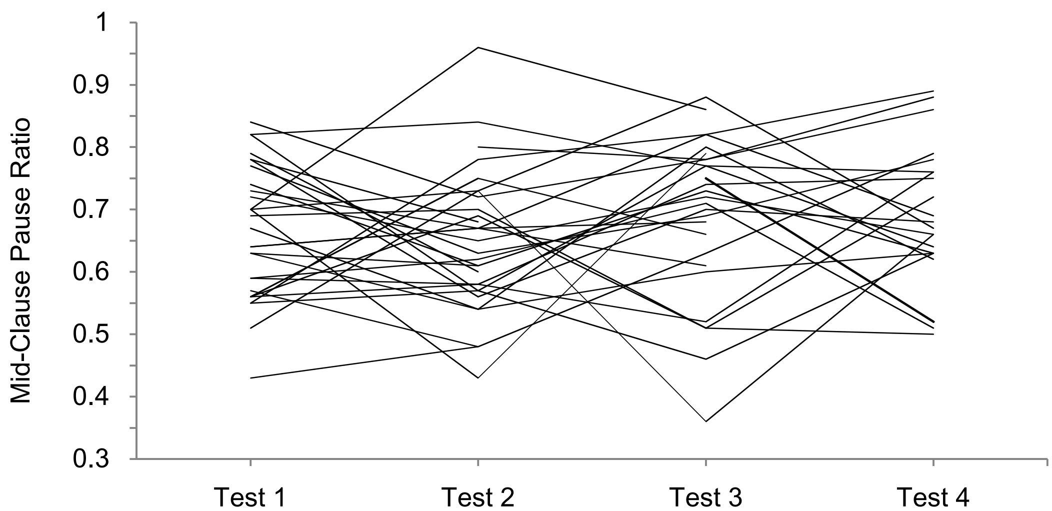

Figure 11. Changes in the mid-clause pause ratio in monologue speaking tests.

In summary, when considering the MLM results, there was significant improvement in the students' pause lengths and proportion of pausing to speaking, however, none of the predictor variables could account for the large amount of unexplained variance. It should be noted that due to the small sample size, statistical power was an issue with these analyses, and different results may have been found with a larger group of students. The implications of these findings are discussed in the next chapter.

Research Question 2: Short-Term Fluency Development

Research question 2 concerned the degree to which students' short-term fluency developed in the 3/2/1 speaking activity. This question was answered by using multi-level linear modeling to determine the growth trajectory of the fluency measures in the 3/2/1 activity across deliveries. Because the construct of fluency is theorized to have more than one component, four separate analyses were run, as improvement in one area of fluency could occur with a decrease in another. In the first model, the mean length of run was the dependent variable, and time was the independent variable. In the second model, the mean length of pause was the dependent variable, and time was the independent variable. In the third model, the articulation rate was the dependent variable, and time was the independent variable. Finally, in the fourth model, the phonation/time ratio was the dependent variable, and time was the independent variable.

To determine whether the fluency measures changed significantly over deliveries in the 3/2/1 speaking activity, three-level MLM models were analyzed. The first analysis conducted for each measure was to inspect the unconditional model, which investigated the development of the measure without any explanatory variables. If the unconditional model shows that there is significant development among the individuals in the study, further analysis is justified. The equation for the unconditional models is:

$$Y_{ti} = \pi_{0ti} + \pi_{1ti}(SPEECH) + e_{ti}$$

In this equation, Y_{ti} is the fluency measure at time t for learner i, π_{0ti} is the initial measure in fluency in the first delivery, π_{1ti} is the growth rate for learners i and e_{ti} is the residual associated with learners i at time t. The level two unconditional model is shown below. In this equation, the initial measure of fluency and the growth rates are random coefficients. The initial measure of fluency is represented by

$$\pi_{0i} = \beta_{00} + r_{0i}$$

$$\pi_{1i} = \beta_{10} + r_{1i}$$

The unconditional model for MLR was run with no predictors and the results are shown in Table 52.

Table 52. *Statistical Results of the Unconditional Model of Mean Length of Run for 3/2/1 Speeches*

	Fixed effects			
	Coefficient	SE	*t*-ratio	*p*
Intercept (mean length of run growth over time) β_{10}	.63	.03	18.05	<.0001
	Random effects			
	Variance	*df*	*χ2*	*p*
Between-test variability (Intercept)	.15	85	220.62	<.0001
Within-test variability	.29			
Reliability (Intercept)		.61		
Intraclass Correlation		.34		

Note. *N* = 32.

On average, the students' mean length of run measures increased over the speeches at .63 of a syllable per delivery, as can be seen in the coefficient for the fixed effects in Table 52. Furthermore, this improvement was significant ($p < .001$). The random effects do show that there was significant variance in student performance in each speech (var (u_1) = .15, $\chi^2(85) = 220.62$, $p < .0001$. The intraclass correlation showed that 34% of the variance in mean length of run in 3/2/1 speeches was attributable to when the test was administered, while 66% was attributable to the students. In summary, there was significant group variation in students' performance across the speeches in the activity, and there also was significant growth in the mean length of run measure. These results can also be seen in Figure 12, which shows a random sample of 20% of the student performances of the 3/2/1 activities. As can be seen from the inspection of this graph, the majority of lines are clearly moving in an upward trend, signaling an increase in the mean length of run.

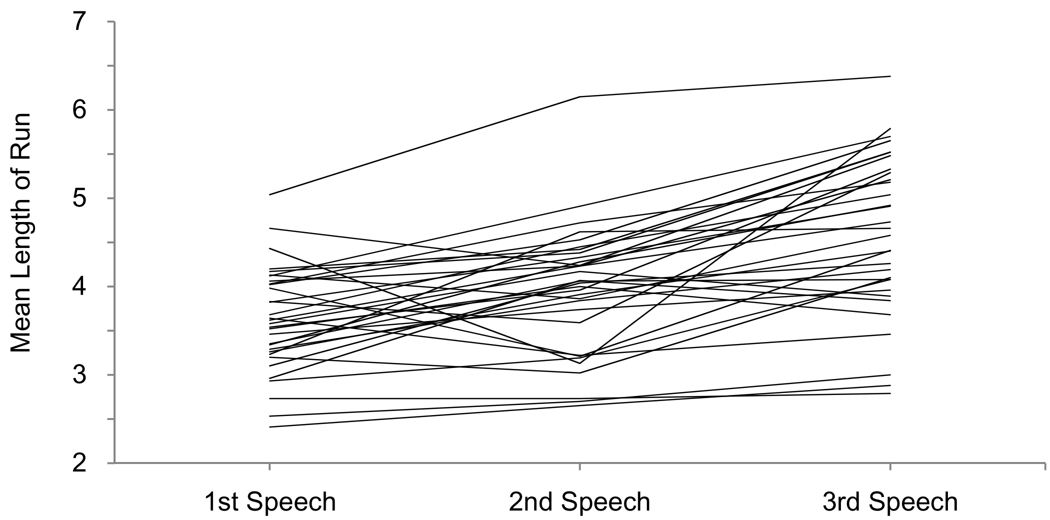

Figure 12. Changes in the mean length of run in 3/2/1 speaking tasks.

The intraclass correlation of .34 indicated that there was a large amount of unexplained variance; thus, each of the independent variables was tested to determine if they could explain any group variance. Given the exploratory nature of this study and the small sample size, each independent variable was run in a separate model, the results of which can be seen in Table 53. An example of the equation for these analyses, in this case, the number of words repeated in all five speaking activities, is:

$$\pi_{1i} = \beta_{10} + \beta_{11}(\text{Rep5}) + r_{1i}$$

Although several of the variables had a statistically significant impact on the improvement of the mean length of run, the size of the coefficient variables in Table 53 show that the effects were negligible, and did not explain any of the individual variance in the ICC. This can be seen by examining the coefficient values in Table 53, which show the impact of the variable measured in syllables. Both of these values are extremely slight, the largest being 0.2 of a syllable for the total number of words used in all five speaking activities.

Table 53. *Estimation of Significant Effects of Predictor Variables on Mean Length of Run*

Fixed effect	Coefficient	*SE*	*t*-ratio	*df*	*p*
R5	.0213	.00282	7.57	420	< .001
Spoken	.0004	.00009	4.50	420	< .001
Form	.0035	.00915	.38	420	.700
Extraversion	.0002	.00039	.56	420	.580
WTC	.0004	.00032	1.41	420	.160
L1 Fluency	-.0001	.00029	-.47	420	.640

Note. R5 = total number of words repeated in all five speaking activities, Spoken = number of tokens spoken in the lesson, Form = number of the types of formulaic language used in the lesson, WTC= willingness to communicate, *N* = 32.

The unconditional model for the mean length of pause was run with no predictors and the results are shown in Table 54.

Table 54. *Statistical Results of the Unconditional Model of Mean Length of Pause for 3/2/1 Speeches*

	Fixed effects			
	Coefficient	SE	*t*-ratio	*p*
Intercept (mean length of run growth over time) β_{10}	-.18	.01	-11.67	<.0001
	Random effects			
	Variance	*df*	*χ2*	*p*
Between-test variability (Intercept)	.007	85	156.00	<.001
Within-test variability	.026			
Reliability (Intercept)		.45		
Intraclass Correlation		.22		

Note. N = 32.

On average, the students' mean length of pause measures decreased over the speeches at .18 of a second per delivery, as can be seen in the coefficient for the fixed effects in Table 54. Furthermore, this improvement was significant ($p < .001$). The random effects do show that there was significant variance in student performance in each speech (var (u_1) = .007, $\chi^2(85) = 156$, $p < .001$. The intraclass correlation showed that 22% of the variance in mean length of pause in 3/2/1 speeches was attributable to when the test was administered, while 78% was attributable to the students. In summary, there was significant group variation in students' performance across the speeches in the activity, and there also was significant development in the mean length of pause measure. These results can also be seen in Figure 13, which shows a random sample of 20% of the student performances of the 3/2/1 activities. As can be seen from the inspection of this graph, the majority of lines are moving in a downward trend, signaling a decrease in the mean length of pause.

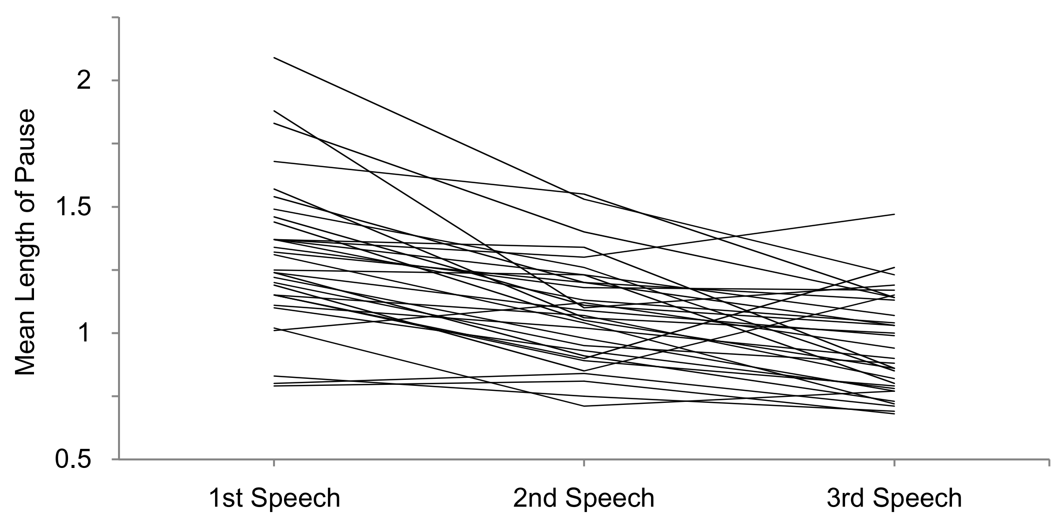

Figure 13. Changes in the mean length of pause in 3/2/1 speaking tasks.

The intraclass correlation of .22 indicated that there was a large amount of unexplained variance. In the same manner as the mean length of run analysis above, each of the performance and individual differences variables were individually examined. Although several variables had a statistically significant impact on the improvement of the average length of pause, the size of the coefficient variables in Table 55 show that the effects were negligible, and did not explain any of the individual variance in the ICC. This can be seen by examining the coefficient values in Table 55, which show the impact of the variable measured in seconds. Both of these values are extremely slight.

Table 55. *Estimation of Significant Effects of Predictor Variables on Mean Pause Length*

Fixed effect	Coefficient	*SE*	*t*-ratio	*df*	*p*
R5	.00485	.00185	2.61	420	.009
Spoken	.00024	.00005	4.60	420	< .001
Formulaic	.00747	.00426	1.76	420	.080
Extraversion	.00009	.00002	-.45	420	.647
WTC	-.00016	.00013	-1.15	420	.250
L1 Fluency	-.00020	.00008	-2.27	420	.024

Note. R5 = total number of words repeated in all five speaking activities, Spoken = number of tokens spoken in the lesson, Form = number of the types of formulaic language used in the lesson, WTC= willingness to communicate, *N* = 32.

The unconditional model for the articulation rate was run with no predictors and the results are shown in Table 56.

Table 56. *Statistical Results of the Unconditional Model of the Articulation Rate for 3/2/1 Speeches*

	Fixed effects			
	Coefficient	SE	*t*-ratio	*p*
Intercept (mean length of run growth over time) β_{10}	9.48	.86	10.97	<.001
	Random effects			
	Variance	*df*	$\chi 2$	*p*
Between-test variability (Intercept)	101.08	85	263.04	<.001
Within-test variability	145.17			
Reliability (Intercept)		.68		
Intraclass Correlation		.41		

Note. N = 32.

On average, the students' articulation rate decreased over the speeches by 9.48 syllables per delivery, as can be seen in the coefficient for the fixed effects in Table 56. Furthermore, this improvement was significant ($p < .001$). The random effects do show that there was significant variance in student performance in each speech (var (u_1) = 101.08, $\chi^2(85) = 263.04$, $p < .001$. The intraclass correlation showed that 41% of the variance in the articulation rate in 3/2/1 speeches was attributable to when the test was administered, while 59% was attributable to the students. In summary, there was significant group variation in students' performance across the speeches in the activity, and there also was significant development in the articulation rate. These results can also be seen in Figure 14, which shows a random sample of 20% of the student performances of the 3/2/1 activities. As can be seen from the inspection of this graph, the majority of lines are clearly moving in an upward trend, signaling an increase in the articulation rate.

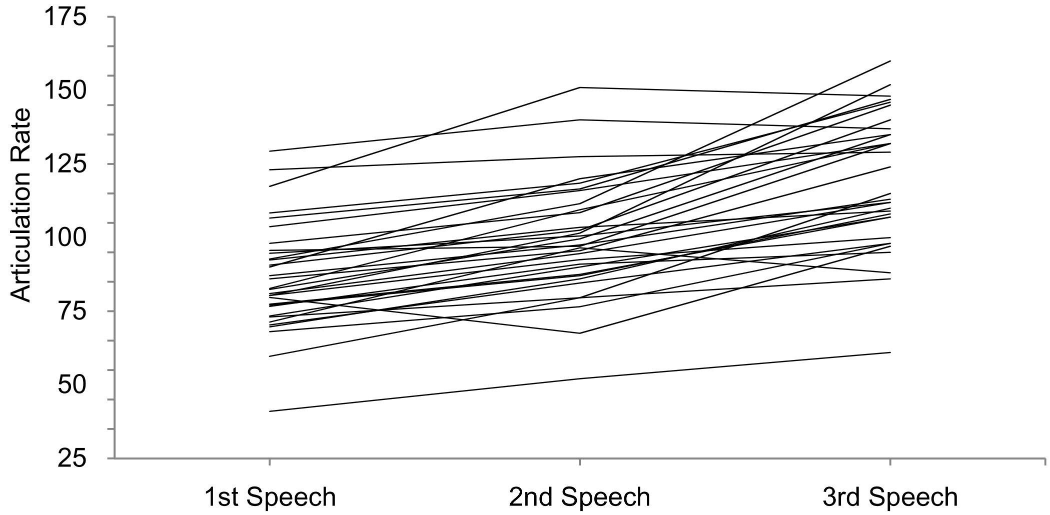

Figure 14. Changes in the articulation rate in 3/2/1 speaking tasks.

The intraclass correlation of .41 indicated that there was a large amount of unexplained variance. In the same manner as the mean length of run analysis above, each of the performance and individual differences variables were individually examined. As can be seen in Table 57, none of the variables had a statistically significant impact on the improvement of the articulation rate.

Table 57. *Estimation of Significant Effects of Predictor Variables on Articulation Rate*

Fixed effect	Coefficient	*SE*	*t*-ratio	*df*	*p*
R5	-.1056	.086	-1.32	420	.219
Spoken	.0002	.002	-.10	420	.919
Formulaic	-.0135	.193	-.07	420	.944
Extraversion	-.0058	.008	-.73	420	.464
WTC	-.0090	.006	-1.51	420	.131
L1 Fluency	.0007	.006	.14	420	.892

Note. R5 = total number of words repeated in all five speaking activities, Spoken = number of tokens spoken in the lesson, Form = number of the types of formulaic language used in the lesson, WTC= willingness to communicate, *N* = 32.

The unconditional model for the phonation/time ratio was run with no predictors and the results are shown in Table 58.

Table 58. *Statistical Results of the Unconditional Model of the Phonation/Time Ratio for 3/2/1 Speeches*

	Fixed effects			
	Coefficient	SE	*t*-ratio	*p*
Intercept (mean length of run growth over time) β_{10}	.06	.003	18.12	<.001
	Random effects			
	Variance	*df*	$\chi2$	*p*
Between-test variability (Intercept)	.0005	85	145.56	<.001
Within-test variability	.002			
Reliability (Intercept)		.41		
Intraclass Correlation		.19		

Note. N = 32.

On average, the students' phonation/time ratio increased over the speeches by .06 per delivery, as can be seen in the coefficient for the fixed effects in Table 58. Furthermore, this improvement was significant ($p < .001$). The random effects do show that there was significant variance in student performance in each speech (var (u_1) = .0005, $\chi^2(85) = 145.56$, $p < .001$. The intraclass correlation showed that 19% of the variance in the phonation/time ratio in 3/2/1 speeches was attributable to when the test was administered, while 81% was attributable to the students. In summary, there was significant group variation in students' performance across the speeches in the activity, and there also was significant development in the phonation/time ratio. These results can also be seen in Figure 15, which shows a random sample of 20% of the student performances of the 3/2/1 activities. As can be seen from the inspection of this graph, the majority of lines are clearly moving in an upward trend, signaling an increase in the phonation/time ratio.

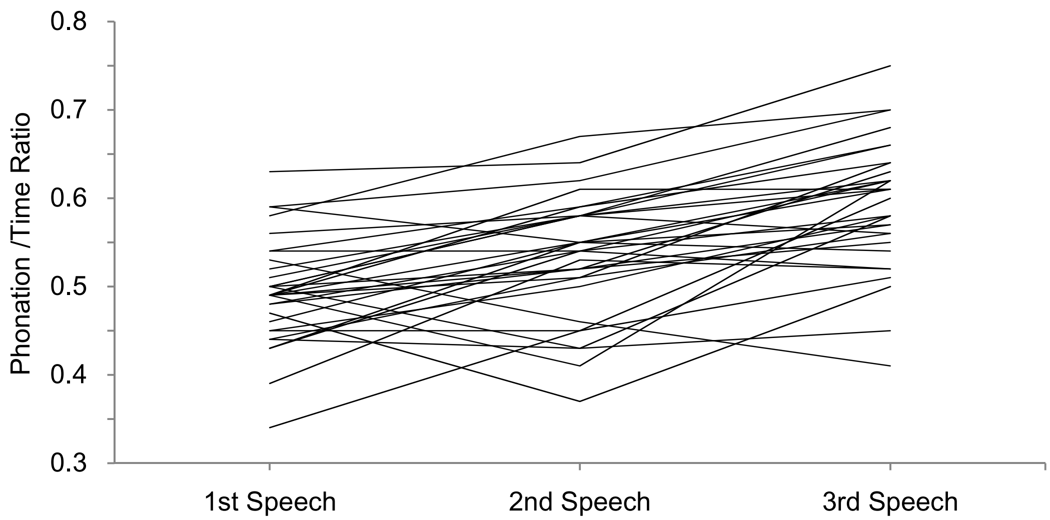

Figure 15. Changes in the phonation/time ratio in 3/2/1 speaking tasks.

The intraclass correlation of .19 indicated that there was a large amount of unexplained variance. In the same manner as the mean length of run analysis above, each of the performance and individual differences variables were individually examined. Although several variables had a statistically significant impact on the improvement of the phonation/time ratio, the size of the coefficient variables in Table 59 show that the effect was negligible, and did not explain any of the individual variance in the ICC. This can be seen by examining the coefficient value in Table 59, which show the impact of the variable measured in syllables. This value is extremely slight, indicating a very small effect.

Table 59. *Estimation of Significant Effects of Predictor Variables on Phonation/Time Ratio*

Fixed effect	Coefficient	*SE*	*t*-ratio	*df*	*p*
R5	.000828	.00049	1.67	420	.095
Spoken	-.000011	.00001	-.87	420	.385
Form	-.001358	.00083	-1.62	420	.106
Extraversion	.000022	.00003	.67	420	.506
WTC	.000062	.00002	2.87	420	.004
L1 Fluency	-.000003	.00002	-.11	420	.910

Note. R5 = total number of words repeated in all five speaking activities, Spoken = number of tokens spoken in the lesson, Form = number of the types of formulaic language used in the lesson, WTC= willingness to communicate, *N* = 32.

Research Question 3: Relationship Among Complexity, Accuracy, and Fluency

The third research question investigated the relationship among complexity, accuracy, and fluency. To date, most of the research concerned with fluency development has not included examinations of how complexity and accuracy might develop in relation to fluency, leading some researchers to claim that the 4/3/2 activity and its variants might harm certain aspects of language development (Boers, 2014; Thai & Boers, 2016). To investigate this question, a bi-variate Pearson correlation analysis (two-tailed) was conducted for each administration of the monologue speaking tests. Table 60 shows that there were no significant correlations among the CAF variables in the first monologue speaking test.

Table 60. *Correlations Among CAF Measures (Monologue Speaking Test 1)*

	1	2	3	4	5
1. Mean length of run	—				
2. Mean length of pause	-.25	—			
3. Accuracy	.18	-.20	—		
4. Syntactic complexity	.28	-.13	.36	—	
5. Lexical complexity	-.07	-.12	-.36	-.20	—

Note. N = 29.

Table 61 shows that there was one positive correlation between the mean length of pause and accuracy in the second monologue speaking test.

Table 61. *Correlations Among CAF Measures (Monologue Speaking Test 2)*

	1	2	3	4	5
1. Mean length of run	—				
2. Mean length of pause	-.36	—			
3. Accuracy	.36	-.52**	—		
4. Syntactic complexity	-.02	-.23	.13	—	
5. Lexical complexity	-.04	-.21	.06	-.30	—

Note. N = 25.
** p < .01.

Table 62 shows that both of the fluency variables, the mean length of run and the mean length of pause, correlated with the accuracy measurement (the ratio of error free clauses to total clauses) in the third monologue speaking test. While not significant, the relationship between the mean length of run and lexical complexity was also stronger in this test.

Table 62. *Correlations Among CAF Measures (Monologue Speaking Test 3)*

	1	2	3	4	5
1. Mean length of run	—				
2. Mean length of pause	-.62**	—			
3. Accuracy	.52**	-.43*	—		
4. Syntactic complexity	.09	.09	-.10	—	
5. Lexical complexity	.24	-.07	.10	-.10	—

Note. N = 30.
** p < .01 , * p < .05.

Finally, Table 63 shows that both of the fluency variables, the mean length of run and the mean length of pause, correlated with the accuracy measurement (the ratio of error-free clauses to total clauses) in the third monologue speaking test. Additionally, the syntactic complexity measurement (the number of clauses per AS-unit) also significantly correlated with a fluency measure (the mean length of run) and the accuracy measure. Finally, the relationship between the mean length of run and lexical complexity remained insignificant, however was of a similar size. This may have been due to proficiency levels, as the counterbalanced design of the test topics meant that differing levels of students took the same test on different administrations of the monologue speaking tests.

Table 63. *Correlations Among CAF Measures (Monologue Speaking Test 4)*

	1	2	3	4	5
1. Mean length of run	—				
2. Mean length of pause	-.52**	—			
3. Accuracy	.58**	-.52**	—		
4. Syntactic complexity	.48**	-.30	.70**	—	
5. Lexical complexity	.26	-.10	.07	-.08	—

Note. n = 28.
** p < .01.

An examination of these tables suggests that over time, student levels of fluency, accuracy, and syntactic complexity became increasingly related to each other. Furthermore, although MTLD did not significantly correlate with the mean length of run, there were indications of a weak relationship between MTLD and the mean length of run in the final two tests. With a larger *n*-size, these relationships might become significant. Similarly, it should be noted that although many of the relationships among the mean length of pause, accuracy, and complexity were not significant, there does appear to be weak relationships in most of the analyses. With a larger *n*-size, some of these relationships might become significant.

Research Question 4: Expert Ratings of Perceived Fluency

Research question 4 investigated the degree to which expert ratings of perceived fluency related to temporal measures of utterance fluency. To investigate this question, a bi-variate correlation analysis (two-tailed) was conducted between the average rating given on each performance with the five fluency measures: the mean length of run, the mean length of pause, the articulation rate, the phonation/time ratio, and the mid-clause pause ratio. The results can be seen in Table 64.

Table 64. *Correlations Among Expert Ratings and Fluency Measures*

	1	2	3	4	5
1. Rating	—				
2. Mean length of run	.35	—			
3. Mean length of pause	-.64**	-.15	—		
4. Articulation rate	.71**	.17	-.35	—	
5. Phonation/time ratio	.53*	.58*	-.74**	.12	—
6. Mid-clause pause ratio.	.32	-.60*	-.14	.38	-.34

Note. $N = 32$.
** $p < .01$, * $p < .05$.

The ratings significantly correlated with three of the five fluency ratings. The strongest relationship can be seen in the articulation rate ($r = .71$, p (two-tailed) $< .01$), which suggests that the raters were focused on the amount of speech produced by the speakers. The ratings also significantly correlated with both the mean length of pause ($r = -.64$, p (two-tailed) $< .01$), and the phonation/time ratio ($r = .53$, p (two-tailed) $< .01$). Based on these results, there appears to be a moderate relationship between the utterance fluency measures, which are objectively measured by analyzing speakers' performances, and subjective ratings of perceived fluency.

Summary

In this chapter I reported on the results of the study designed to answer the research questions in Chapter 2. Regarding the first question, the EFL students' levels of spoken fluency did not develop over one academic semester, regardless of the in-class performance or individual difference variables that were examined. The second question showed that speaking fluency developed significantly across the three deliveries of the 3/2/1 speaking activity. This development however, was seen in all students, regardless of their performance before the activity, or their individual difference variables.

The third research question concerned the relationships among the complexity, accuracy, and fluency (CAF) measures across the semester. The results indicated that over the semester, each of the CAF variables, with the exception of lexical complexity, became more closely related to one another. Finally, research question 4 investigated the degree to which expert ratings of perceived fluency were related to utterance fluency measurements. The results indicated that apart from mean length of run, the expert ratings had a significant and moderate correlation with the remaining utterance fluency measurements.

In the next chapter, I summarize the results for each research question, and then discuss the findings in detail with reference to existing research. The implications of the findings for researchers and teachers are also discussed.

CHAPTER 6

DISCUSSION

In this chapter, I restate the research questions and summarize the findings of the study. Following this, I interpret the results of each research question. Finally, the research and pedagogical implications with regards to second language fluency development are discussed.

Research Question 1: Long-Term Fluency Development

Research question 1 investigated the degree to which EFL students' levels of spoken fluency developed over one academic semester. Previous studies have shown that speaking fluency can develop over relatively short periods of time (as little as ten weeks), for students studying in ESL contexts (De Jong & Perfetti, 2011; Mora & Valls-Ferrer, 2012) and study abroad contexts (Freed et al., 2004; Segalowitz & Freed, 2004, Towell et al., 1996), but there have been few longitudinal studies of fluency development in EFL contexts, meaning that it is unclear whether students who have limited exposure to English can benefit in similar ways. Furthermore, most of the ESL studies cited in this study were focused on higher proficiency, more mature students, most of whom were using the second language in higher education settings. It remains to be seen whether non-English major students studying in their first language can make significant improvements in compulsory foreign language classes that are not the main focus of their studies.

The most widely used criterion to determine whether fluency has developed is to compare the development of two variables: the mean length of run and the mean length of pause. If either of these measures has improved whilst the other remains stable, an argument can be made that students have become more fluent speakers. In this study, four commonly used measures of fluency - the mean length of run, the mean length of pause, the phonation/time ratio, and the articulation rate - were tracked over four monologue speaking tests conducted over an academic semester. In addition, the ratio of mid- to end-clause pauses was also looked at to investigate whether students were able to reduce the amount of lexical or syntactic encoding needed to produce speech. Furthermore, I also sought to discover if fluency development covaried with in-class performance or individual difference variables. Therefore, the relationship among fluency development and several in-class performance variables - participation, the number of words repeated, and the amount of formulaic language used—and individual difference variables—willingness to communicate, extraversion, and L1 fluency—were also investigated.

Summary of the Results for Long-Term Fluency Development

The results showed that two of the fluency measures, the mean length of pause and the phonation/time ratio, made small, but significant improvements over the course of the semester. Both of these variables are based around the amount of pausing and hesitation in an oral performance. However, no significant amounts of variation were observed in the mean length of run, the articulation rate, and the mid-clause pause ratio, which fluctuated for most learners over the course of the study. Therefore, the developments in spoken fluency were positive, but very slight, especially considering that the mean length of run and articulation rate are the fluency variables calculated from the amount of speech produced by a learner.

As only two of the fluency measures showed significant development, additional analyses were only conducted for the mean length of pause and the phonation/time ratio. For both measures, the Intra-class Correlation Coefficients indicated that a large amount of variance, .47 for the mean length of pause, and .57 for the phonation/time ratio, remained to be explained by variables related to the individual speakers. However, none of the predictor variables were able to significantly account for any of this variance. Therefore, no further analyses were conducted with regards to this research question. Finally, it should be noted, that due to the small sample size, statistical power was an issue with this study, and results may have differed with a larger sample size.

Interpretation of the Results for Long-Term Fluency Development

There are several insights to be found, given the lack of research on fluency development in EFL settings. First, students appear to become more fluent speakers even with limited opportunities to use English. Second, this development did not appear to be directly related to any of the in-class performance measures or the individual difference measures included in the study (extraversion, willingness to communicate, and L1 fluency).

These results are somewhat similar to the findings of other fluency development studies conducted in ESL settings. The closest similarities can be found with De Jong and Perfetti's study (2011), in which their largest experimental group of ten students made similar gains. A rough comparison of the total change from the first to the last test also shows a similar reduction in pause length and the phonation/time ratio, both for the students in this study and those in the repetition condition in De Jong and Perfetti's study. In this study, the average gain between the first and the second monologue test was .19 seconds, compared with an average gain of .15 seconds for the students in the De Jong and Perfetti study. Similarly, the average gain in the phonation/time ratio in this study was 4.14%, compared with 3.88% in the De Jong and Perfetti study.

With regard to the remaining measures, the mean length of run and the articulation rate, a comparison can also be made between the participants in this study and those in the first repetition condition of De Jong and Perfetti's study in that no significant gains were made in these measures. However, while the participants in the De Jong and Perfetti study made small non-significant increases in these measures of .16 syllables for mean length of run and 9.88 syllables for the articulation rate, there were small negative values seen with these measures in this study of -.02 syllables for mean length of run and -2.06 syllables for the articulation rate. Finally, it should be noted that participants in the second repetition condition in De Jong and Perfetti's study ($n = 5$) made progress in the opposite manner: their mean length of run improved significantly while the mean length of pause remained stable.

Looking at other longitudinal studies of fluency development, the results of this study can be interpreted positively, as the learners did make some small gains in fluency. When compared with the lack of gains seen in the EFL comparison groups that have been used in research looking at the relationship between participating in study abroad programs and fluency development (Freed et al., 2004; Mora & Valls-Ferrer, 2012; Segalowitz & Freed, 2004), the participants in this study appeared to make better gains than those in the control group students. However, it must be stressed that there is not a lot of information given about the 'at home' participants in these studies, and it seems unlikely that they took part in fluency development practice activities or extensive speaking time in their classes, as most descriptions imply that these classes were general English classes with few opportunities for students to speak.

Compared with the De Jong and Perfetti study, there are several factors that might have influenced the lack of fluency development. First, the students in prior study were speaking to themselves in front of a computer terminal. This study was designed to determine whether fluency developed in regular classroom activities with students interacting with each other. This condition however, might have affected the fluency measures, as on some occasions, especially when speaking with friends, the speaker and listener interacted for brief periods of time. Second, the participants in the De Jong and Perfetti study completed the 4/3/2 activity three times in a language class, speaking on topics that were unrelated to their studies. Therefore, it is possible that the novelty of the activity could have affected these performances positively. In contrast, the students in this class completed the activity every week, and were speaking about topics they had been studying. While the students did not appear bored or fatigued by the activity, its regularity might have resulted in some performances that were lower than students' actual abilities, and failed therefore, to drive fluency development that could be transferred to performance in the speaking monologue tests.

While there are some similarities between these results and those found in the De Jong and Perfetti (2011) study, there are also some important differences. De Jong and Perfetti found a positive correlation between the amount of lexical overlap, specifically words used in all three deliveries of the 4/3/2 speaking activity, with gains in two of the monologue speaking test measures; the mean length of pause and the phonation/time ratio. These correlations were significant and of moderate size. In this study, the amount of lexical overlap was not related to any gains in the monologue tests made by the students. However, it should be noted that De Jong and Perfetti included their control group in their correlation analysis. The students in the control group were given different topics for each delivery of their 4/3/2 speeches and consequently were not able to repeat similar language. On average, the participants in the control group only repeated 5.3 words across all three deliveries in the 4/3/2 speaking task, compared with 23.9, and 30.8 words for the two repetition conditions. The students in the control group also did not improve in their speaking fluency measures. Therefore, this reported relationship between repetition and fluency development might have been artificially strengthened by including students who spoke under different conditions in the same analysis.

In comparison, all of the participants in this study repeated far more words than the students in the De Jong and Perfetti study; an average range of 24.8 to 48.3 over each of the eight 3/2/1 speaking activities. Table 65 shows the average number of words repeated in each of the 3/2/1 speaking activity. Several observations can be made from these data. First, it is clear that the amount of repetition appears related to proficiency level, as the students from the higher level classes are much more frequent at the top of the chart than are those from the lower level classes. Second, a comparison with the words repeated in the 3/2/1 activity and the words repeated over the five class activities shows that students who repeated a lot of words in the 3/2/1 activity were not always repeating words that they had used earlier in the lesson. This partially explains why the in-class measure of words repeated across five activities did not account for any of the variance in fluency development, although it should be noted that repetition limited to the 3/2/1 speeches also did not significantly account for any variance.

Finally, while this point should be interpreted with caution, most of the students seemed to improve in either the mean length of pause or the mean length of run. In fact, all but three students (highlighted in grey in Table 65) made improvements between their first and final test, suggesting that if repetition does enable fluency development, it might do so in different aspects of fluency for different students. The research methodology of comparing mean scores for students on their mean length of run and mean length of pause could therefore, obscure some of the gains made by students who improved in one area but not in the other. It should be emphasized, however, that the gain figures in Table 65 were derived by subtracting the score of the students' first monologue speaking test from their last, and do not therefore, take the variation over the four tests into account, as did the MLM model reported in the previous chapter. Nevertheless, these findings point to areas that future studies of fluency could focus on.

Table 65. *Amount of In-Class Repetition and Gain Scores on the Monologue Speaking Tests*

Students	R3/2/1	R5	Average mean length of pause gain	Average mean length of run gain
Int #1	48.29	23.25	-.01	-.02
High Int #1	48.14	19.17	.68	-.16
High Int #2	45.63	14.50	-.17	-.84
Int #2	45.17	28.43	.37	.01
High Int #3	43.88	15.88	.56	-.31
High Int #4	40.67	20.75	.99	.44
Int #3	40.33	16.00	-.79	-.32
High Beg #1	39.00	11.14	1.04	.77
High Int #5	38.88	16.83	-.07	-.17
Int #4	38.50	14.33	.34	.13
Low Int #1	37.63	18.67	.26	-.25
Low Int #2	36.43	9.75	.72	-.06
High Int #6	36.33	21.14	.51	-1.00
Low Int #3	36.00	18.13	-.67	-.04
Int #5	34.33	17.17	-.09	-.25
Low Int #4	33.00	12.88	-.10	-.01
High Beg #2	33.00	16.67	.00	.79
Low Int #5	31.63	14.40	.48	-.16
High Beg #3	31.63	18.29	.58	-.07
Low Int #6	31.17	23.00	.15	-.24
High Beg #4	31.14	18.38	-1.89	-.22
Int #6	30.50	14.50	.15	-.78
Int #7	29.86	19.50	-.22	-.48
High Int #7	29.67	8.25	-.32	-.82
Low Int #7	29.60	13.63	-.29	.08
High Beg #5	29.57	10.14	.01	.11
High Beg #6	29.14	12.88	-.20	-.04
High Beg #7	29.00	15.57	.08	.37
Low Int #8	28.00	13.50	-1.39	.19
High Beg #8	25.13	13.71	.34	-.02
Int #8	24.88	12.40	-.89	-1.05
High Beg #9	24.80	17.50	-.06	1.04

Note. High int = High intermediate, Int = intermediate, Low int = low intermediate, High Beg = high beginning, R3/2/1 = number of words repeated in 3/2/1 speaking activity, R5 = number of words repeated in all five speaking activities, *N* = 32.

Based on the rough approximation of gains that can be seen in Table 65, it appeared that many students were improving on their mean length of run or their mean length of pause. Given that these two fluency variables are closely related, a post-hoc analysis was run to determine whether more of the variance in student development of the mean length of pause could be explained when the mean length of run was taken into account. The results, reported in Table 66, show that participants who had higher levels in their mean length of run had significantly shorter pause rates. When taking this factor into account, the unaccounted variance from the ICC dropped from .47 to .35, meaning that initial levels of mean length of run explained more than 10% of the variance in student performance. Given the post-hoc nature of this analysis, however, the results showed be interpreted cautiously and warrant further investigation in future studies.

Table 66. *Estimation of Baseline for Initial Status and Growth on Mean length of pause*

Fixed effect	Coefficient	*SE*	*t*-ratio	*df*	*p*
MLR	-.131	.052	-2.50	78	.014
Time	-.006	.027	-2.25	78	.027

Note. MLR = mean length of run, *N* = 32.

In addition to the lexical overlap measure, none of the in-class performance variables explained any of the variance in either the mean length of pause or the phonation/time ratio. It appears that the participants were unable to transfer the short-term fluency gains they made in class activities to other speaking tasks that were focused on different topics. This is somewhat surprising given that opportunities for practice and repetition have been theorized to aid the development of speaking fluency (De Jong & Perfetti, 2011).

Another explanation for the lack of fluency development might be the students' lower levels of oral proficiency. The bulk of research published on longitudinal fluency development has focused on high-intermediate or advanced level students (De Jong & Perfetti, 2011; Lennon, 1990; Tavakoli et al., 2016; Towell et al., 1996). The findings of one study on task repetition has also suggested that language proficiency had significant correlations of a moderate size when participants were asked to repeat a narrative task after a one-week interval (Mojavezi, 2013). The gain scores for students arranged by class level in Table 67 suggest similar findings, as on average, only the high intermediate students improved their fluency in both the mean length of run and pause measures. In contrast, the remaining classes, intermediate, low intermediate, and high beginning, only improved their mean length of pause, and made slight losses in the mean length of run. There are however, large standard deviations that show considerable variation in the gain scores. Furthermore, it should be stressed that the figures do not reflect variation over the four monologue tests.

Table 67. *Means and Standard Deviations of Mean Gains in the Mean Length of Pause and the Mean Length of Run*

Level	Average mean length of pause gain	Average mean length of run gain
High intermediate[a]	.31 (.50)	-.41 (.51)
Intermediate[b]	-.14 (.48)	-.35 (.41)
Low intermediate[c]	-.08 (.69)	-.07 (.15)
High beginning[d]	-.03 (.79)	-.16 (.44)

Note. [a]n = 7, [b]n = 8, [c]n = 8, [d]n = 9.

However, based on these findings, two further analyses were conducted to see if proficiency covaried with fluency development. The initial model was run with the students' scores on the placement TOEIC test; however, all analyses fell short of significance, with the closest p value being .07 for the mean length of pause. Therefore, given the small sample size and the exploratory nature of the study, I decided to separate the students into two groups and assign the students in the two lowest classes a dummy code of 0, and the students in the two highest classes a dummy code of one. The multilevel model analyses of the four monologue speaking tests were rerun, and a significant finding was found for the higher proficiency group for the mean length of pause, reported in Table 68, although it should be emphasized that while the finding is significant, the coefficient of -.098 shows that students in the higher proficiency group were able to significantly decrease their mean length of pause about a tenth of a second more on each administration of the monologue speaking tests than those in the lower group.

Table 68. *Estimation of Significant Effects of Predictor Variables on Mean Length of Pause*

Fixed effect	Coefficient	*SE*	*t*-ratio	*df*	*p*
Proficiency	-.098	.049	-2.03	78	.046

Note. Proficiency = Proficiency grouping (0 = classes high beginner and low intermediate classes, 1 = intermediate and high beginner classes), *N* = 32.

This model was then rerun with each in-class and individual difference predictor variable. When rounded, significant results were found for the average amount of words repeated across the five speaking activities. The results can be seen in Table 69.

Table 69. *Estimation of Significant Effects of Predictor Variables on Mean Length of Pause*

Fixed effect	Coefficient	*SE*	*t*-ratio	*df*	*p*
Proficiency	-.113	.048	-2.34	77	.022
Repeated	.007	.003	1.97	77	.052

Note. Proficiency = Proficiency grouping (0 = classes high beginner and low intermediate classes, 1 = intermediate and high beginner classes), Repeated = average number of words repeated in all five speaking activities, *N* = 32.

As there was a very small effect for repetition, the analysis was rerun, substituting the average number of words repeated in the five speaking activities for the average number of words repeated in the 3/2/1 speaking activities. As shown in Table 70, the effect was slightly stronger, although still very small, with the coefficient at .009 of a second, although it was in a positive direction, meaning that the students who used more repetition decreased their pausing on the four monologue speaking tests less than their peers who did not repeat as much language. However, in addition to the small scale of these significant effects, it should be stressed that the post-hoc nature of these analyses should be interpreted with caution and investigated in more detail in further research. Proficiency level does, however, appear to play some role in the effect that fluency activities have on student development.

Table 70. *Estimation of Significant Effects of Predictor Variables on Mean Length of Pause*

Fixed effect	Coefficient	*SE*	*t*-ratio	*df*	*p*
Proficiency	-1.150	.051	-2.95	77	.004
R321	.009	.003	2.93	77	.004

Note. Proficiency = Proficiency grouping (0 = classes high beginner and low intermediate classes, 1 = intermediate and high beginner classes), R321 = average number of words repeated in 3/2/1 speaking activities, *N* = 32.

The amount of time that students were able to use English each week could also explain the very small gains made by the group. In total, the students spent approximately 4.5 hours using English per week, consisting of three English classes: the discussion class focused on in the study, a presentation class, and either a writing or an e-learning class. Both students and teachers of these other classes reported that there was very little time for speaking activities in these other English classes. This finding accords with other findings from other fluency development studies. In the study abroad research, the only foreign language group shown to develop greater speaking fluency was an immersion group who reported using even more of the L2 outside of class than the study abroad group who were living in the target culture (Freed, et al., 2004).

Furthermore, the two studies of fluency instruction that have demonstrated that gains from practice activities can transfer to other tasks were conducted in ESL settings (De Jong & Perfetti, 2011; Tavakoli et al., 2016). In both of these studies, the participants were enrolled in intensive EAP programs. Therefore, it might be necessary for learners to engage in much more language practice outside of speaking classes than the students in this study were able to. Issues related to motivation might also be relevant here; all of the students in this study were enrolled in a compulsory English course in their first year of university, whereas the students in the above studies had invested time and money to study English in a foreign country. It might be the case that because students in ESL contexts have invested time and money to study abroad, they are more motivated to learn and use the target language.

However, while the in-class performance variables did not significantly account for development in speaking fluency, they did have significant relationships with students' performance in the monologue speaking tests. A bi-variate Pearson correlation analysis (two-tailed) was conducted between the in-class performance measures and the four fluency variables. The results can be seen in Table 71.

Table 71. *Correlations Among In-Class Performance and Monologue Test Measures*

	Repeated Words	Spoken	Formulaic
Mean length of run			
Monologue test 1[a]	.55**	.68**	.30
Monologue test 2[b]	.53**	.52**	.34
Monologue test 3[c]	.56**	.60**	.23
Monologue test 4[d]	.64**	.63**	.11
Mean length of pause			
Monologue test 1[a]	-.60**	-.63**	-.39*
Monologue test 2[b]	-.46*	-.71**	-.61**
Monologue test 3[c]	-.45*	-.46*	-.07
Monologue test 4[d]	-.54**	-.62**	-.48**
Articulation rate			
Monologue test 1[a]	.60**	.54**	.20
Monologue test 2[b]	.52**	.54**	.16
Monologue test 3[c]	.52**	.49**	.30
Monologue test 4[d]	.42*	.47*	-.09
Phonation/time ratio			
Monologue test 1[a]	.55**	.68**	.44*
Monologue test 2[b]	.37	.52**	.57**
Monologue test 3[c]	.47*	.51**	.07
Monologue test 4[d]	.54**	.60**	.38

Note: [a]N = 29, [b]N = 25, [c]N = 30, [d]N = 28.
** p < .01, * p < .05.

As can be seen in Table 71, the speakers who exhibited higher levels of performance on the in-class measures also spoke more fluently during the monologue speaking tests, as there are significant moderate correlations among many of the variables. The monologue test measures related to the quantity of speech produced, i.e. the mean length of run and the articulation rate, were significantly correlated with both the average number of words that students repeated over the five activities in each lesson and the average number of words spoken during each lesson. The pausing and hesitation measures, the mean length of pause and the phonation/time ratio, also significantly correlated with most of the formulaic phrase use measures. The figures in

Table 71 suggest that not only did the students who spoke often in class outperform their peers on the monologue speaking tests, but also the students who repeated more of the same language, although no causal relationships can be ascertained from correlation analyses.

As none of the individual difference variables had any significant impact upon fluency development, follow-up analyses were carried out to determine if they had any relationship with the fluency measures on the monologue speaking tests. A bi-variate Pearson correlation analysis (two-tailed) was conducted between the individual difference measures and the four fluency variables. However, there were no significant relationships among the individual differences and the fluency measures, as can be seen in Table 72.

While it is surprising that neither levels of WTC or extraversion were related to levels of speaking fluency, one possible interpretation is related to the unique nature of this particular class. As the largest class in the study only had nine students, it was possible for the teacher to monitor student performance at all times, which meant that even less willing speakers spoke for the duration of the time during the monologue tests; therefore, the volition and/or disposition to speak might not have played a role in student performances. Another possibility is that the test time of 2 minutes was too short; therefore, it might be necessary to extend this time in order to see the relationship among these variables and student performance. Finally, levels of L1 fluency were not shown to have any relation to the fluency measures in these tests. This finding contradicts previous research that has shown significant relationships between these two measures (De Jong et al., 2013; Kahng, 2014). It could, however, be that the speaking task might have been too difficult for students to speak with their usual level of L2 fluency, given that only one minute of planning time was given to students before they had to begin speaking. As is discussed below, most of the students spoke with higher levels of fluency during all three rounds of the 3/2/1 speaking activity. Another possibility is that given their relatively low level of oral proficiency, there was not enough variation in their English speaking fluency to have a clear relationship with L1 speaking fluency.

Table 72. *Correlations Among Individual Difference and Monologue Test Measures*

	Willingness to Communicate	Extraversion	L1 Fluency
Mean length of run			
Monologue test 1[a]	.36	.08	.12
Monologue test 2[b]	.09	-.09	-.05
Monologue test 3[c]	.03	-.14	-.12
Monologue test 4[d]	.29	-.02	.03
Mean length of pause			
Monologue test 1[a]	.13	.22	.35
Monologue test 2[b]	.03	.23	.17
Monologue test 3[c]	.25	.15	.15
Monologue test 4[d]	-.02	.16	.01
Articulation rate			
Monologue test 1[a]	.00	.27	.09
Monologue test 2[b]	-.13	-.32	.02
Monologue test 3[c]	-.12	-.31	-.08
Monologue test 4[d]	-.05	.04	-.21
Phonation/time ratio			
Monologue test 1[a]	.09	-.06	-.23
Monologue test 2[b]	.13	-.10	-.24
Monologue test 3[c]	-.14	-.09	-.13
Monologue test 4[d]	.11	-.07	.05

Note: [a]N = 29, [b]N = 25, [c]N = 30, [d]N = 28.

Finally, before moving on to a discussion of the results for short-term fluency development, a comparison is made between students' fluency measurements on the monologue speaking tests and the three deliveries of the 3/2/1 speaking activity. The purpose of this comparison is to ascertain whether students spoke at a similar rate in the tests as they did in the class activities. In Table 73, the lowest average mean length of run value is highlighted. For 21 of the students, the lowest value was obtained from their monologue test performance, whilst for the remaining 11 students, the lowest value came from their performance in the first delivery of the 3/2/1 speaking activity.

Despite these differences, in many cases the mean length of run measure in both the monologue test and the first three-minute speech is very similar, although it should be emphasized that the monologue test was only 90 seconds long, which complicates the comparison, as the mean length of run increases with longer stretches of speech. Nevertheless, it can clearly be seen that all of the students had much higher rates on the 2- and 1-minute speeches in the 3/2/1 speaking activity. This result suggests that students found it more challenging to produce speech with a minimum of planning time, even though the topics were based around common themes. In contrast, the pre-reading and speaking activities conducted before the 3/2/1 speaking task might have primed students to speak at or above their usual level of fluency.

It should also be noted that fluency is context-dependent, and given that students were engaged in meaningful discussions with an interlocutor in the short-term fluency speaking activities, it could be the case that some students spoke at slower rates to enhance the communicative effectiveness of their message. Although the students were instructed not to interact with each other during the recordings of the monologue tests, the students were addressing a peer, which could have influenced the fluency with which they spoke. In some instances, therefore, a student's level of utterance fluency may not be indicative of their actual level of cognitive fluency.

Table 73. *Mean Length of Run Measures on the Monologue Test and the 3/2/1 Speaking Activity*

Students	Average monologue test mean length of run	Average three minute mean length of run	Average two minute mean length of run	Average one minute mean length of run
High Int #1	3.89	3.98	4.44	5.11
High Int #2	3.21	4.64	5.03	6.02
High Int #3	3.39	3.14	3.52	4.18
High Int #4	5.06	4.76	5.42	6.53
High Int #5	3.08	3.50	3.62	4.42
High Int #6	4.96	4.15	4.68	5.77
High Int #7	2.46	2.63	2.84	3.09
Int #1	2.59	3.05	3.39	3.85
Int #2	3.19	3.71	4.52	6.03
Int #3	3.24	3.97	4.20	5.28
Int #4	4.24	3.93	4.40	5.52
Int #5	3.27	2.96	3.01	3.68
Int #6	4.70	5.17	5.79	6.82
Int #7	4.27	4.40	4.69	5.88
Int #8	3.06	3.59	4.03	4.88
Low Int #1	3.47	3.35	4.00	4.32
Low Int #2	3.66	3.45	3.61	4.22
Low Int #3	3.72	3.73	4.12	4.58
Low Int #4	3.33	3.96	4.19	4.99
Low Int #5	3.81	4.20	4.94	5.20
Low Int #6	3.11	3.46	4.15	4.88
Low Int #7	3.32	3.69	3.91	4.81
Low Int #8	3.32	3.73	4.19	5.07
High Beg #1	2.38	2.39	2.61	3.05
High Beg #2	2.85	2.94	3.24	3.56
High Beg #3	2.35	2.70	3.28	3.79
High Beg #4	3.35	3.20	3.47	3.74
High Beg #5	3.49	3.55	4.10	4.93
High Beg #6	3.40	3.56	4.04	4.82
High Beg #7	3.43	3.11	3.26	3.72
High Beg #8	3.97	3.14	3.17	3.94
High Beg #9	3.57	3.29	3.82	4.86

Note. High int = High intermediate, Int = intermediate, Low int = low intermediate, High Beg = high beginning.

Finally, it should be noted that this study was conducted in the second semester of the academic year, after all students had successfully completed the first semester of the discussion course in addition to two other English courses; presentation and writing for the two highest proficiency classes, and presentation and e-learning for the two lowest proficiency classes. It is possible that larger long-term fluency gains could have been seen if the study was conducted in the first semester, as after graduating high school and entering university, the students would arguably have had lower levels of speaking fluency. A final possibility is that the time period of the study, 1 semester, is simply too short to observe significant changes in utterance fluency. If the study were conducted over a longer period of time, some of the small improvements may have become more pronounced. As the raters were able to distinguish between relatively small differences in utterance fluency, perceived fluency ratings might also improve over longer periods of time.

Research Question 2: Short-Term Fluency Development

Research question 2 investigated the degree to which EFL students' levels of spoken fluency developed over three deliveries of the 3/2/1 speaking task. Previous studies have shown the activity to be beneficial in terms of developing speaking fluency for students studying in ESL contexts (Arevart & Nation, 1991; De Jong & Perfetti, 2011), but no researchers have examined the use of this activity in EFL contexts, with the exception of Thai and Boers (2016), who investigated the activity with high school students in Vietnam. Furthermore, except for the De Jong and Perfetti study (2011), there have been no longitudinal studies of the use of this activity, and some degree of the positive findings could be attributable to the novelty of the activity. The impact of implementing the 3/2/1 speaking task in Japanese university contexts is therefore, unclear.

As with the long-term fluency development research question, the mean length of run and the mean length of pause were compared to determine if there was an increase in short-term fluency. The four commonly used measures of fluency—the mean length of run, the mean length of pause, the phonation/time ratio, and the articulation rate—were tracked over eight 3/2/1 speaking tasks that were conducted over the course of an academic semester. Furthermore, this study also sought to discover if short-term fluency development covaried with in-class performance or individual difference variables. Therefore, the relationship among short-term fluency development and several in-class performance variables (participation, the number of words repeated, and the amount of formulaic language used), in addition to individual difference variables (willingness to communicate, extraversion, and L1 fluency) was also investigated.

Summary of the Results for Short-Term Fluency Development

The results showed that on all of the four fluency measures, the mean length of run, the mean length of pause, the articulation rate, and the phonation/time ratio, the students made significant improvements in the 3/2/1 speaking activity. Moreover, the Intra-class Correlation Coefficients revealed that a fairly large portion of the variance remained to be explained by individual factors. These coefficients ranged from .21 to .45. However, even though some of the predictor variables, specifically the number of repeated words and the total number of words spoken, were significant in the model, closer analysis revealed that they did not significantly explain any of this remaining variance, and overall, had very little impact upon overall student performance.

Interpretation of the Results for Short-Term Fluency Development

These results are consistent with other researchers who have looked at the 4/3/2 (Arevart & Nation, De Jong & Perfetti, 2011) and the 3/2/1 (Thai & Boers, 2016) speaking activities in that fluency significantly improved across the three deliveries. A comparison with the mean scores across the deliveries with the repetition groups in De Jong and Perfetti's study (2011) shows that the fluency measurements appear stronger than those found in this study (see Table 74). De Jong and Perfetti's study was chosen for comparison, as it is the only study which used a 4/3/2 speaking task with the same students on several different occasions. In general, when comparing the average improvement from the first to the third delivery, this comparison supports the utility of the shorter 3/2/1 speaking activity for lower level students, as they were also able to improve their short-term fluency in a similar manner to the higher intermediate level students in the De Jong and Perfetti study, although many students found it difficult to continue speaking for the full three minutes. In general, the students made lengthy pauses in their final statements of most of their deliveries, which suggests that a longer period of time could have been too challenging, although these pauses do appear shorter for the higher level students. The excerpts below, taken from Appendix N, are from the final statement made in the first three-minute delivery, and show that there were quite lengthy pauses for all students, excepting Maya, who had just switched to a new topic.

Excerpt 1: Sho (high beginner student)

next (3.72) and (2.17) i want to learn (3.87) foreign culture (1.98)

Excerpt 2: Saori (low intermediate student)

and (.27) Nara and Kyoto have many traditional (.86) culture (.32) so (1.45) they (.60) they want to go to (1.71) there (1.49) and (2.06)

Excerpt 3: Maya (intermediate student)

so it is interesting for foreigners (.60)

and difficult thing is J-movie because (.84)

Excerpt 4: Sachi (high intermediate student)

then (1.71) not (.44) only pop culture but (.41) traditional culture is (2.06) we want to (1.18) foreigners to know because.

It should also be pointed out that when comparing the average improvement gains seen at the bottom of Table 74, the students in this study appeared to make greater gains than the students in the De Jong and Perfetti study. This advantage could be related to the shorter length of their activity, a total of 6 minutes, as opposed to 9 minutes in the De Jong and Perfetti study. Furthermore, as no statistical analyses were conducted, it is unclear whether the differences are statistically significant.

Table 74. *Comparison of Mean Scores of Fluency Measures with De Jong and Perfetti (2011)*

	This study	De Jong & Perfetti
First delivery		
Mean length of run	3.59	4.62
Mean length of pause	1.27	.96
Articulation rate	180.64	216.19
Phonation/time ratio	.49	.56
Second delivery		
Mean length of run	3.99	4.86
Mean length of pause	1.08	.90
Articulation rate	187.84	218.68
Phonation/time ratio	.54	.59
Third delivery		
Mean length of run	4.73	5.24
Mean length of pause	.91	.84
Articulation rate	200.52	226.59
Phonation/time ratio	.61	.61
Average improvement from first to third delivery		
Mean length of run	1.14	.62
Mean length of pause	-.36	-.12
Articulation rate	19.88	10.4
Phonation/time ratio	.12	.05

Unlike previous 4/3/2 and 3/2/1 research, I sought to determine if any in-class performance (the number of repeated words, the total number of words spoken in the lesson, and the total number of functional language types used in the lesson) or individual difference (extraversion, willingness to communicate, and L1 fluency), covaried with short-term fluency development. The results of the MLM analysis showed that both the number of words repeated in all five activities, and the total quantity of words, were significant predictors of most of the fluency measures. However, the strength of these predictors was very weak and indicated that the measures had a trivial impact upon development across the speeches. Due to this finding, it was decided to take a closer analysis of the relationship among the measures for each of the eight 3/2/1 speaking activities, with the aim of identifying whether there were any conditions under which the relationships were stronger in order to follow up in future research.

To identify the lessons in which the two significant in-class measures were more closely related to fluency development, correlation analyses were conducted among these variables (the number of words used in all five lessons, and the total number of tokens spoken) and the fluency gains for each lesson. The results can be seen in Table 75.

The results showed that the number of words repeated in all five activities was only significantly related to mean length of run gains in four of the eight lessons: lessons 1, 4, 6, and 8. To investigate this result further, a paired-samples *t*-test was conducted to determine whether the lessons with significant correlations in the above analysis had higher levels of gain in the mean length of run than the lessons that did not have a significant relationship with the amount of repetition. To conduct the analysis, each lesson with a significant relationship was paired with the chronologically closest lesson that did not have a significant relationship. Therefore, lessons 1, 4, 6, and 8 were correspondingly paired with lessons 2, 3, 5, and 7, and paired samples t-tests were conducted. If a student was absent from one of the lessons, their data were eliminated, leaving 180 from a possible 256 performances. Before running the analysis, the assumption of normality was checked, and eight outliers were found, three from the correlated lessons and five from the not correlated lessons. These scores were reduced to the mean plus two standard deviations (Field, 2009), reducing the mean in the significantly related lessons from 1.28 syllables per run to 1.24, and the mean in the not significantly related lesson from 1.0 syllables per run to .88. The descriptive statistics can be seen in Table 76.

Table 75. *Correlations Among In-Class Performance Measures and Fluency Gains in the 3/2/1 Activities*

	Mean length of run gain	Mean length of pause gain	Articulation rate gain	Phonation/time ratio gain
Lesson 1[a]				
R5	.51**	.19	.20	.10
Spoken	.37*	.50**	.11	-.20
Lesson 2[b]				
R5	.28	.17	-.15	.15
Spoken	.22	.60**	-.21	-.26
Lesson 3[c]				
R5	-.04	.28	.10	-.19
Spoken	.21	.60**	.06	-.35
Lesson 4[d]				
R5	.44*	.04	-.17	.40*
Spoken	.04	.30	-.12	-.09
Lesson 5[e]				
R5	.01	.18	.04	-.13
Spoken	.17	.56	.04	-.37
Lesson 6[f]				
R5	.56**	.29	.05	.14
Spoken	.46*	.39*	.15	-.04
Lesson 7[g]				
R5	.33	.53**	-.17	.02
Spoken	.43*	.43*	.10	-.17
Lesson 8[h]				
R5	.50**	.16	.40*	.06
Spoken	.19	.32	.18	-.13

Note. [a]$N = 30$, [b]$N = 29$, [c]$N = 27$, [d]$N = 27$, [e]$N = 26$, [f]$N = 28$, [g]$N = 30$, [h]$N = 30$.
** $p < .01$, * $p < .05$.

Table 76. *Descriptive Statistics for Gains in Mean Length of Run in Lessons Related to Amount of Repetition and Lessons Not Related to Amount of Repetition*

	n	*M*	*SE*	*SD*	*Skewness*	*SES*	Kurtosis	*SEK*
Sig. w/Rep.	95	1.24	.08	.80	.21	.25	-.51	.49
Not Sig. w/Rep.	95	.88	.06	.59	.48	.25	.30	.49

The results indicated that the average mean length of run gain in the correlated lessons (*M* = 1.24, *SD* = .80) was significantly greater than the average in the lessons that were not (*M* = .88, *SD* = .59), *t*(94) = 3.62, *p* = .00. The standardized effect size index *d*, was .37, indicating a medium effect (Field, 2009). The 95% confidence interval for the mean difference between the gain scores was .16 to .57.

As the mean length of run gains were significantly higher in the lessons where the gains were positively correlated with the amount of repetition, one possible interpretation is that for repetition to have a positive effect on short-term fluency development, it might have been necessary for students to exceed a certain gain threshold of approximately 1 syllable or greater over the course of the three deliveries. This implies that on the occasions where the students were willing and/or able to 'push' their output, they were able to use the repetition to contribute to a gain in fluency. In contrast, the situations where repetition was not related to gains in the mean length of run, the majority of students repeated the same information with the same vocabulary, but might not have been willing or able to produce language at a faster rate. It should be emphasized however, that this was a post-hoc analysis based on the study's results. Further research is necessary to determine whether a threshold of gain in the mean length of run must be exceeded before students can use repetition to help them improve their speaking fluency.

It should also be noted that the in-class variables sometimes positively correlated with the gains in the mean length of pause, that is, they were associated with less reduction in pausing. In five of the lessons, the total number of words spoken correlated positively with changes in pauses, meaning that students who used more words were also pausing for longer. This could be related to the proficiency level of the students as they might have used these pauses to plan longer stretches of speech. This finding again suggests the difficulty of ascertaining which variables affect fluency development, as students can attend to speech production at the expense of pause rates, while others might be able to attend to both.

Finally, as with the monologue speaking tests, it should be noted that fluency is context-dependent, and given that students were engaged in meaningful discussions with an interlocutor in the short-term fluency speaking activities, it could be the case that some students spoke at slower rates to enhance the communicative effectiveness of their message. In some instances, therefore, a student's level of utterance fluency may not be indicative of their actual level of cognitive fluency.

Research Question 3: Relationships Among Complexity, Accuracy, and Fluency

Research question 3 concerned the relationships among complexity, accuracy, and fluency in the monologue speaking tests. One common criticism of some longitudinal studies of fluency development is that by and large, researchers have not reported on whether complexity and accuracy development are impacted by improvements in fluency (Boers, 2014; Thai & Boers, 2016). To date, one experimental study was designed to address the issue of accuracy in fluency development activities. Thai and Boers (2016) compared one group that completed a 3/2/1 repeated narrative activity with a control group that completed an unpressured 2/2/2 repeated narrative activity, although it should be emphasized that the study was cross-sectional, not longitudinal. The students in the 3/2/1conditon made significant improvements in fluency, whereas the students in the 2/2/2 condition made smaller but significant gains in accuracy. While not supported by empirical evidence, the authors claimed that fluency activities might cause learners to automatize incorrect language forms, as repeating the same strings of words across deliveries might lead learners to internalize inaccurate language chunks. This study therefore, represents an attempt to trace the relationships among complexity, accuracy, and fluency (CAF) to determine whether fluency developed at the expense of complexity and accuracy.

Summary of the Results for Relationships Among Complexity, Accuracy, and Fluency

In the initial monologue test, none of the CAF measures were related to each other. In the second test, there was a significant correlation between the mean length of pause and the accuracy measure, the error-free clause ratio, (r = -.52, p < .01). In the third test, this relationship remained at a similar level (r = -.43, p < .05), and there were two additional significant correlations; the error-free clause ratio also significantly correlated with the other fluency measure, the mean length or run, (r = .52, p < .01), and one between the mean length of run and the mean length of pause

($r = -.62$, $p = .00$). In the final test, these relationships remained significant. The mean length of run correlated with the mean length of pause ($r = -.52$, $p < .01$) and the error-free clause ratio ($r = .58$, $p < .01$). The error-free clause ratio also retained a significant correlation with the mean length of pause ($r = -.52$, $p < .01$). Additionally, syntactic complexity, measured by the number of clauses per AS-unit, correlated with two of the measures; the mean length of run ($r = .48$, $p = .01$), and the error-free clause ratio ($r = .70$, $p = .01$). These results suggest that the CAF measures became more related to one another over time.

While lexical complexity did not correlate significantly with the other CAF measures, there was a small, nonsignificant relationship between this variable and the mean length of run in the last two tests, .24 for monologue test 3, and .26 for monologue test 4. With a larger sample size, this relationship may have become stronger, although it is unclear why the relationship was much smaller in the earlier tests, -.07 in the first monologue test, and -.04 in the second. This fluctuation could have been due to the counterbalanced rotation of the test topics, which meant that different students spoke on different topics for each of the tests.

Finally, it should be emphasized that apart from a small but significant increase in the mean length of pause, none of the variation in the remaining CAF measures were significant. Therefore, although the relationship among the variables became stronger, there was no significant improvement in student performance.

Interpretation of the Results for Relationships Among Complexity, Accuracy, and Fluency

An analysis of the results of the monologue tests suggest that the learners paid attention to more aspects of their language use over time. Initially, no CAF variables had any significant relationship with one another, yet by the end of the study, there were significant relationships among measures from each of the CAF constructs. While there has not been a great deal of research that has investigated how the relationships among CAF variables change over time, especially with EFL students in instructed settings, one of the most influential theories related to CAF research is that of Skehan (1998; 2014), who argued that learners often prioritize one or two of the CAF constructs due to attentional capacity limitations. One possible interpretation of the results is that the students were increasingly able to attend to more aspects of their language while they spoke. This could be due to their growing familiarity with the task; as time went on, they became more able to focus on a wider range of language features when they were speaking.

The results of this study also provide some insight into the criticism that task repetition activities with time-pressure might cause learners to automatize their errors (Boers, 2014; Thai & Boers, 2016). The results showed that there was no significant change in accuracy or complexity over the course of the semester, meaning that although learners engaged in the activity on a weekly basis, syntactic errors did not increase as time went on. To investigate this issue in more detail, the syntactic errors on the monologue tests were coded into the four categories of error used in a study of accuracy and instructed learners in an ESL context (Tonkyn, 2012). The four categories were: (a) noun phrase errors (related to incorrect use of number/case, or determiners), (b) verb phrase errors (related to incorrect inflection or auxiliary verb use, (c) syntactical errors related to incorrect word order or the omission or misuse of essential clause elements, and (d) lexical errors (related to the incorrect use of an open class word, i.e. a lexical verb, noun, or adjective, or adverb). The number of words per error could then be calculated to determine if any kinds of errors were more prevalent than others. This ratio was taken from a study conducted by Tonkyn (2012), who was able to discern gains in ESL student levels of accuracy over a 10-week period. The results can be seen in Table 77, which shows that there was a fair level of fluctuation for the error types in the different tests. It is important to note that when interpreting the data in this chart, higher values indicate higher levels of accuracy, as students were able to produce more language without making an error of the type in question.

Table 77. *Means and Standard Deviations of Error Types in Monologue Speaking Tests*

Error type	Test 1[a]	Test 2[b]	Test 3[c]	Test 4[d]
Words/NP error	348.58 (231.12)	394.20 (245.22)	387.00 (254.18)	315.63 (179.84)
Words/VP error	584.26 (367.13)	529.15 (354.04)	478.13 (339.97)	482.49 (282.16)
Words/syntax error	343.62 (194.39)	424.10 (370.80)	414.35 (390.15)	438.66 (297.29)
Words/lexical error	703.10 (369.29)	829.60 (500.85)	784.23 (410.06)	728.14 (322.16)

Note. [a]*N* = 29, [b]*N* = 25, [c]*N* = 30, [d]*N* = 28, NP = noun phrase, VP = verb phrase.

To investigate the argument that learners might automatize their errors through participation in the 3/2/1 activity (Boers, 2014; Thai & Boers, 2016), a correlation analysis was run among the error types for each monologue test and two measures of repetition: the average amount of lexical overlap in the 3/2/1 speaking activity, and the average amount of lexical overlap across the five activities in each speaking class. A further analysis was made between gain in words per error type (noun phrase, verb phrase, syntactical, and lexical) and the measures of lexical repetition. Gain scores were calculated by subtracting the number of words per error type from the first test taken to the last; therefore, these results are a rough approximation and should be interpreted with caution. While these analyses cannot reveal causal relationships between repetition and levels of accuracy, a significant negative relationship between the number of words per error and the amount of repetition would suggest that learners who repeat more of the same words also use less accurate language, and be a good candidate for further research. The results in Table 78, however, show that there were no significant relationships in a negative direction. On the contrary, there were several positive correlations in both the tests and the gains, which suggest that the students who used more repetition were also able to use longer stretches of accurate language. It is also notable, that on three of the four tests, there were significant correlations of a moderate size between the amount of repetition and the number of words per lexical error. This finding suggests that students who were able to repeat more words had a better knowledge of vocabulary, which might have allowed them to confidently use the same language when repeating their speeches.

Furthermore, an inspection of some of the errors across the 3/2/1 activities also provide insight into the issue of whether learners automatize their errors. Rather than repeat each error verbatim in the subsequent speech, there seemed to be considerable variation in word order and usage. This variation can be seen in excerpts from learner transcripts in Appendix N. For ease of reading, repetitions, self-corrections, and reformulations have been removed, and the errors discussed have been underlined.

Table 78. *Correlations Among Mean Use of Repetition and Words Per Error Type*

	Words/NP error	Words/VP error	Words/syntax error	Words/lexical error
Test 1[a]				
R321	.03	.34	-.19	.20
R5	-.20	.35	-.17	.32
Test 2[b]				
R321	.42*	.39	.48*	.60**
R5	.42*	.33	.19	.54**
Test 3[c]				
R321	.02	.41*	.31	.48**
R5	.08	.51*	.11	.59**
Test 4[d]				
R321	-.03	.17	-.14	.47*
R5	-.14	.33	-.18	.57**
Gains[e]				
R321	-.05	-.16	.35*	.37*
R5	.06	.38	.09	.27

Note. [a]*N* = 29, [b]*N* = 25, [c]*N* = 30, [d]*N* = 28, [e]*N* = 32, R321 = average number of repeated words in 3/2/1 activities, R5 = average number of words repeated in the five classroom activities.
** $p < .01$, * $p < .05$.

In excerpt 5, the speaker reduces the content of the initial statement in each delivery and does not repeat the mistake. These problems were not restricted to lower-proficiency students, and be seen across the upper three proficiency levels as well, where some errors were corrected, others remained, and new errors arose.

Excerpt 5: Sho (high beginner student)

3 minutes: i like manga very much so i want to introduce Japanese manga <u>for</u> foreign people *[syntax error: incorrect use of preposition "for"]*

2 minutes: i like manga very much so i want to discuss my favorite manga

1 minute: i want to discuss my favorite manga

This result can be seen in excerpt 6, where the speaker is able to correct the errors in the verb and noun phrases, but uses an incorrect lexical item.

Excerpt 6: Saori (low intermediate level student)

3 minutes: we <u>have not</u> <u>chance</u> to wear kimono *[verb phrase error: incorrect inflection, noun phrase error: missing determiner]*

2 minutes: but we don't have <u>chance</u> to <u>put on</u> kimono *[noun phrase error: missing determiner, lexical error: incorrect use of 'put on']*

1 minute: kimono is very stylish but we don't have a chance to <u>put on</u> *[lexical error: incorrect use of 'put on']*

Sometimes erroneous statements were repeated verbatim, however, in excerpt 7, after repeating an incorrect phrase in the first two deliveries, Maya is able to use the correct form in her final speech.

Excerpt 7: Maya (intermediate level student)

3 minutes: we talked about which culture do you want to introduce to <u>a foreigner friends</u> *[noun phrase error: incorrect determiner or plural use, syntax error: noun form used instead of adjective 'foreign']*

2 minutes: we talked about which culture do you want to introduce to a foreigner friends *[noun phrase error: incorrect determiner or plural use, syntax error: noun form used instead of adjective 'foreign']*

1 minute: we talked about pop culture or traditional culture which culture do you want to introduce to foreigners *[no errors]*

Students did not always correct their incorrect statements, however. In excerpt 8, there are several errors in each reformulation of the phrase. While each iteration features incorrect usage of prepositions, in the first two phrases, the student correctly uses the definite article before '*textbook*'. However, in the final delivery, she does not, and also incorrectly adds a determiner to '*pop culture*,' and error that is not seen in the first two speeches:

Excerpt 8: Sachi (high intermediate level student)

3 minutes: at first time we think pop culture is important because as in the textbook it says eighteen billion dollars as like pop culture like manga and anime or something like that [changes topic] *[5 syntax errors: incorrect use of noun following the phrase 'at first', 2 incorrect uses of preposition, unnecessary insertion of subject after textbook, and missing verb phrase following 'eighteen billion dollars']*

2 minutes: it get profits like in the textbook it shows it's profit was eighteen billion dollars or something like that so maybe pop culture is really moving

economy *[verb phrase error: subject verb agreement, noun phrase error: missing determiner]*

1 minute: the pop culture is good because in textbook it says eighteen billion dollars are earned from profit *[2 noun phrase errors: unnecessary and missing determiners]*

A close inspection of the transcripts in Appendix N shows numerous errors repeated across all three deliveries; however, there are also many examples of instances where students correct themselves, or make new errors with language that was previously accurate. It appears, therefore, that the students' language was fairly dynamic across the performances.

Although it is beyond the scope of this study to identify the exact reasoning for this behavior, one explanation is the students' proficiency levels; their lack of proficiency might have had an effect on their ability to remember and reuse the language from a prior delivery. Furthermore, when listening to these audio recordings, the students occasionally used rising intonation followed by a pause, which can be interpreted as them checking to see if their utterance has been understood. This behavior might signal that the students themselves were unsure if whether what they said was correct, due to their lack of experience using English both outside and inside of the classroom. Many students remarked that they had never discussed most of the topics covered by the text in English classes.

In summary, while there were no significant changes in levels of accuracy or complexity, there was no evidence to support the claims that time-pressured fluency development activities result in learners proceduralizing incorrect language forms. Furthermore, the students who repeated more language made fewer lexical errors during the tests, which suggests that repetition is associated with more accurate performance rather than less.

Research Question 4: Relationship Among the Expert Ratings and Utterance Fluency Measures

Research question 4 concerned the degree to which expert ratings of the students' speech were related to utterance fluency measures. While previous studies (De Jong & Perfetti, 2011; Tavakoli et al., 2016) have shown significant improvement in utterance fluency measures within a short period of time, one criticism is that these gains are very slight on paper. Specifically, the significant improvement in the mean length of run in the De Jong and Perfetti study was an average of .86 syllables in the mean length of run. While the same statistic was a higher 1.67 in the Tavakoli et al. study, it is unclear whether such small changes could be detected by human raters, questioning the validity of these significant improvements in utterance fluency.

Therefore, I decided to ask a group of experts to rate the learners' performances to determine the relationship between the objective utterance fluency measures and the more subjective perceived fluency utterances. These raters listened to two monologue test performances from eight students. In terms of utterance fluency, one performance was the student's highest mean length of run, and the other performance was their lowest. On average, the difference between these high and low performances were approximately 1.5 syllables, which is comparable to the gain found in the Tavakoli et al. study (2016). After playing the raters the recordings in random order, the scores were analyzed using a many-faceted Rasch analysis to control for rater severity and to investigate the validity of the rating scale.

Summary of the Results for Relationship Among the Expert Ratings and Utterance Fluency Measures

Preliminary analysis of the rating scale used to determine perceived fluency showed that the raters did not make full use of the lowest two categories, as only one rater made use of the lowest rating. The two lowest descriptors were therefore collapsed into one category, and the resultant four-point scale was shown to function satisfactorily. Investigation of fit statistics showed that the raters all fit the requirements of the Rasch model, however one student misfit the model due to higher ratings on both performances from one rater. Separation indices showed that there were significant differences in the students' highest and lowest performances, indicating that, in general, the raters were able to distinguish the faster performances from the slower.

A correlation analysis was then run between the expert ratings and student levels of complexity, accuracy, and fluency on the monologue tests from which the recordings were taken. The results showed that three of the five fluency measures significantly correlated with the expert ratings, the mean length of pause, the articulation rate, and the phonation/time ratio. The two strongest relationships were between the mean length of pause (r = -.64, p [two-tailed] < .01), and the articulation rate ($r = .71$, p [two-tailed] $< .01$). The size of the relationship was slightly lower for the phonation/time ratio ($r = -.51$, p [two-tailed] $< .05$).

These findings indicate that despite the small differences in the mean length of run between performances, the raters were able to perceive which speeches were the most fluent. This finding has positive implications for the relatively small gains in utterance fluency seen in studies such as De Jong and Perfetti (2011) and Tavakoli et al. (2016), as these changes appear to be perceptible to the naked ear.

Interpretation of the Results for Relationship Among the Expert Ratings and Utterance Fluency Measures

The results indicate that expert listeners are indeed able to detect relatively small differences in utterance fluency, as determined by the mean length of run measure that was used to select the learners' highest and lowest performances. Interestingly, the mean length of run and the mid-clause pause ratio were the only fluency variables that did not significantly correlate with the ratings. This finding could be due to the fact that these measures are based on relatively specific averages that the raters were unable to pay attention to. However, the results do suggest that even quite modest differences in utterance fluency measures can be detected by the human ear.

When comparing utterance and perceived fluency measures, Segalowitz (2010) cautions that there is a circularity involved in arguing that fluency can be considered both an impression of a speaker's proficiency level (perceived fluency) and an objective measure of oral output, and that fluency measures are particularly sensitive to the elicitation task used to measure them. Therefore, while the raters' judgments largely agreed with the objective measurements, it should also be cautioned that this relationship might not be found if different tasks were used. Given that there was a large degree of variation in student performance and no clear development trajectory, it could be that the tests that were given higher ratings were simply those that were the easiest for students to speak on.

Despite this concern, the results indicated that expert raters can be sensitive to very slight differences in levels of utterance fluency. Significant correlations of a moderate size were found to exist between fluency measure related to both the quantity of speech produced and the amount of time spent to deliver the message. Given the small sample size and the relative homogeneity of student proficiency, these results should be explored in different contexts to determine the degree to which they can be generalized.

Theoretical Implications of the Results

The findings of this study have several implications for researchers interested in the development of second language speaking. First, it appears possible that EFL learners in instructed settings can make small gains in fluency over an academic semester, similar to the findings in instructed ESL settings (De Jong & Perfetti, 2011). While these gains do appear relatively minor, there were significant correlations of a moderate size between ratings of perceived fluency and the objective fluency measures, indicating that these changes can be detected by the naked ear. It should be emphasized that this study was conducted over a relatively short period of time;

therefore, it is possible that EFL learners may be able to continue with this improvement, leading to more noticeable improvement in perceived fluency, if they continue to engage in fluency development activities over longer periods of time.

A striking finding in this study is that utterance fluency development did not appear to covary with performance in classroom activities, or by the individual difference variables that were included in the analyses. This might have been due to the design of the study, in particular, the fact that students spoke to one another in pairs. This feature is in contrast to previous studies where students have given monologue speeches with no interlocutor present (De Jong & Perfetti, 2011; Tavakoli et al., 2016). While this method of data collection does have more ecological validity for those interested in fluency development in communicative language classroom settings, there might have been too much distraction from having interlocutors present. In some instances, it may even have been the case that students adjusted their speaking speed to accommodate their peers with lower levels of language proficiency. There does however, need to be further research conducted into this area to ascertain the effect that the context of the speaking setting has on fluency development.

Given that none of the individual difference variables influenced long- or short-term fluency development, the question arises whether the levels of utterance fluency measured in this study were influenced by the context in which the data was gathered. It could be the case, that similar to some of the studies of the relationship between extraversion and L2 proficiency (Dewaele & Furnham, 2000), the classroom setting was too informal for the individual difference variables to affect levels of utterance fluency. As the study was conducted in the second semester of the year and the students knew each other well, their familiarity with one another may have dampened any influence that extraversion might have had on the degree to which they spoke. Furthermore, given the small class size and the ability of the instructor to monitor each group, willingness to communicate may also have played less of a role in the volition with which students spoke. Follow-up studies to compare the relationship between the context, for example formal and informal settings, or with familiar and unfamiliar peers, and these individual differences might show us more about how an individual's personality can affect the degree of fluency with which they speak.

Second, while there was some minor development in the mean length of pause and the phonation/time ratio measures, the results of the study also show a great deal of variability in the students' levels of utterance fluency. These differences were detected by the expert raters, who did not know that they had listened to two recordings of each speaker. One source of this variability could have been the prompts for each monologue test. There has been research showing that similar test prompts can elicit different results in fluency and lexical measures (De Jong & Vercellotti, 2016). However, the results of that study indicated significant differences in the speaking prompts, whereas in this study, the differences seem to arise from the learners themselves, as there were no significant differences found when comparing the topics. Evidence for this possibility can further be seen in Table 79, which shows the topic with the number of students who gave their best and worst performance in terms of their mean length of run. The one student who gave his highest performance on two different prompts is included twice in the highest column. This variation in performance suggests that, especially for learners at lower levels of proficiency, the language sample obtained from the prompt could be highly dependent upon the learner's personal relation to it. Therefore, it might be desirable to obtain several fluency measures on a variety of topics when investigating fluency development with lower proficiency students, or those studying in foreign language contexts where they have few opportunities to use English outside of the classroom.

As mentioned above, there may also be a strong interlocutor effect when measuring the utterance fluency of learners interacting with their peers. For this reason, studies investigating fluency development in instructed settings should also include fluency measures obtained when students are not addressing an interlocutor. While it is unclear whether the higher level students did slow down to make themselves more comprehensible to their less able peers, there were several instances in the monologue speaking tests where the listeners interacted with the speaker. Follow-up studies should be conducted to determine the relationship between utterance fluency measures on speeches with and without interlocutors.

Table 79. *Topics and Students' Highest and Lowest Mean Length of Run Measures*

Topic	Number of students with highest MLR	Number of students with lowest MLR
Travel abroad vs. Japan	5	1
City vs. country	6	2
Youtube vs. movies	3	1
TV vs. computer games	3	4
Playing vs. watching sport	4	1
Playing vs. listening to music	0	4
Summer vs. winter	2	3
Birthdays vs. new year	4	3
Eating at home vs. restaurants	5	6
Time alone vs. with friends	1	7

Note. MLR = mean length of run.

Pedagogical Implications of the Results

In terms of short-term fluency development, the results support the findings that repeated time-pressure activities can effectively develop students' abilities in this area. This study included a repetition variable that had not been used in prior fluency developments studies; the amount of repetition used in activities prior to the time-pressured speaking task. The results showed that in certain cases, there was a positive relationship between the number of words repeated across the five speaking activities, which does seem to support the integration of time-pressured speaking tasks with topics that are focused on in the learners' courses, as they might be able to build their oral fluency through the meaningful repetition of previously discussed ideas.

However, as with the monologue speaking tests, the in-class performance measures and the individual difference variables did not appear to have any significant effects on the gains made between speech deliveries. Several points can be made here. First, the effects that repetition has on fluency development in the short term could be obscured by two conflicting issues related to task repetition. Follow-up analyses suggested there was only a significant relationship between repetition and short-term fluency gains when the mean length of run gain was higher than just over 1 syllable between the first and third delivery. It could be the case that only when learners were engaged with the topic and the lesson, they were able to use repetition to increase their speaking fluency. Again, this could be related to the selection of topics, as the lessons in which repetition and fluency had a significant relationship with each other were relatively simple and related directly to students' lives, such as studying abroad, celebrities and role-models, public rules, and gender stereotypes. In contrast, the topics where repetition did not significantly correlate with fluency gains could arguably be seen as slightly more abstract, and therefore more difficult for the students, i.e. being fashionable, information and the Internet, traditional culture vs. popular culture, and the death penalty. Again, this possibility points to the need for more investigation into the effects of topic familiarity and/or engagement with measures of speaking fluency.

Furthermore, post-hoc analyses of the results suggested that learners with higher levels of proficiency made more improvements, both in the mean length of run and pause, than their peers with lower levels of proficiency. These findings suggest that it may be learners may benefit more from fluency instruction once a certain threshold of proficiency has been gained. For this reason, theories of communicative language teaching such as Nation's four strands (2001), which argues for a balanced approach between meaning-focused input and output, form-focused, and fluency activities, may suit learners of a higher level of proficiency. In contrast, lower level learners may benefit more from controlled and semi-controlled practice activities that allow them to engage in meaningful interactions in a more structured setting. Such types of activities may allow learners to develop the levels of accuracy and lexical knowledge needed to deliver an extended monologue in a fluent manner. It should however, be emphasized that follow-up studies are required in order to determine the relationship between proficiency level and the gains made from courses that include fluency development activities.

In terms of the formulaic language phrases covered in the course, there were no significant relationships found between its use and fluency gains made in either the short- or the long-term. This might have been due to the nature of the phrases, as many of them are used to manage turn-taking and group interactions, rather than to structure a monologue. To aid learners with the ability to chunk their output into longer fluent runs, it might therefore, be more effective to cover formulaic language directly concerned with the organization of monologue speeches, that is, transitions between points, ways to change the topic, abandon an idea, or make reference a previously mentioned comment.

Summary

In this chapter, I have discussed and summarized the results for each of the four research questions in the study. Following that, I attempted to interpret the findings in reference to the results from similar studies and with follow-up analyses on questions brought to light by the results. In general, while some similarities could be seen with regard to longitudinal fluency development studies, the lack of research into factors other than study context made it difficult to interpret some of the findings of the in-class performance and individual predictor variables.

In Chapter 7 I briefly summarize the main findings of the study. I then discuss the limitations of the study, before making suggestions for future research in second language fluency development studies.

CHAPTER 7

CONCLUSION

In this chapter, I begin with a brief summary of the main findings of the study, and then consider some of the limitations of these findings. Next, I make several suggestions for future research before stating my final conclusion about the role of fluency development activities in EFL instructed settings.

Summary of the Findings

This study was designed to shed light on the development of speaking fluency in instructed EFL settings. As the need to speak English fluently is a goal for many foreign language learners, it is important for language teachers to know more about fluency development and the factors that can affect it. While much of the pedagogical literature on teaching speaking recommends fluency development activities (Gatbonton & Segalowitz, 1988; Nation & Newton, 2009), there has been little investigation into how fluency instruction affects learners in foreign language instructed settings. In fact, it is not even well-known whether learners in instructed settings can effectively develop their speaking fluency, especially those learners who have limited access to English. Research into the effects of study abroad on fluency development have used ‘at home’ groups as control comparisons (Freed, et al., 2004; Mora & Valls-Ferrer, 2012; Segalowitz & Freed, 2004), and the results from these studies show that in all cases, there were no significant gains in speaking fluency made by the students enrolled in traditional language courses at their university.

Second, if EFL learners can develop their speaking fluency in instructed settings, it is important to identify the specific characteristics of the learners who make the largest gains. Several researchers have argued that repetition (De Jong & Perfetti, 2011) and formulaic language (Wood, 2010) are key variables that influence fluency development trajectories. Furthermore, individual difference variables that have been argued to play a key role in language learning such as willingness to communicate could also have an influence on the size of learners' gains.

The final motivation for this study was to learn more about what happens to student levels of complexity and accuracy as fluency develops. While some researchers have argued that fluency activities can also aid accuracy and complexity (Arevart & Nation, 2011), time pressured fluency development activities such as the 4/3/2 speaking tasks and its variants have received criticisms by researchers concerned that learners might automatize their errors through repeated inaccurate language (Boers, 2014; Thai & Boers, 2016). Therefore, it needs to be determined whether fluency activities can be harmful to student development.

In this study, I sought to address these issues by tracing the fluency development of 32 learners from four intact discussion-skills classes. These classes were relatively unique in a Japanese university setting, as there were a maximum of nine students per class. To determine the degree of fluency development that learners were able to attain, a research design similar to that used by De Jong and Perfetti (2011) was employed. Student performance on a fluency development activity, which was a reduced 3/2/1 version of the 4/3/2 speaking task, in addition to performance in two previous activities were collected over one academic semester. In addition, the students participated in four monologue speaking tests to determine if the short-term gains made in the 3/2/1 speaking activities transferred to other speaking tasks. In comparison with the De Jong and Perfetti study, this study included more data collection points, specifically five more fluency development activities, as well as two additional speaking activities that preceded them, and had more assessments, as four monologue tests were administered as opposed to three.

The first research question focused on the long-term fluency gains made in these monologue tests. The results showed that apart from small significant improvements in the mean length of pause and the phonation/time ratio, there was no significant growth in fluency development. However, while the improvements were very small, the possibility remains that over periods of time longer than the length of this study, these small changes might become more pronounced. The expert raters were able to distinguish between the relatively small differences in student performances in this study, therefore, the possibility remains that if a longer time frame were examined, there might be more striking changes in both utterance and perceived fluency. However, none of the in-class or individual predictor variables explained any of the variance in the growth of the two significant variables. Further post hoc analysis of the results showed that the positive growth in the mean length of pause might covary with students' mean length of run, and their proficiency level, although follow-up studies need to be conducted to learn more about whether these findings are representative of fluency development for EFL learners, or simply chance relationships that were found in the post-hoc analyses.

The second research question focused on the degree to which students made short-term fluency gains when participating in the 3/2/1 speaking task. The results showed that students made significant improvements in all four fluency measurements: the mean length of run, the mean length of pause, the articulation rate, and the phonation/time ratio. However, despite some significant findings for two of the measures, the amount of words repeated in the lesson's five activities, and the total number of words spoken in the lesson, the predictor variables did not appear to have any relationship with these fluency gains. Further post-hoc analyses showed that repetition had significant correlations of a moderate size with the mean length of run gains made in four of the eight lessons. A follow-up paired samples *t*-test showed that in these four lessons the mean length of run was significantly higher than it was in the four lessons where repetition did not significantly correlate with the mean length of run gains. It appears, therefore, that for repetition to have a positive effect on short-term fluency development, a threshold of an approximate 1.1 syllable gain was needed to be made across the three speeches. Given the post-hoc nature in which these findings were discovered, it is necessary however, to conduct follow-up studies to determine whether this is the case.

The third research question focused on the effects that fluency development might have on complexity and accuracy. The results showed that over time, the learners' complexity, accuracy, and fluency measures became more positively related to each other. These findings dispute the claims that learners automatize inaccurate language forms through participation in time-pressured speaking activities. In fact, a follow-up analysis revealed that in three of the four monologue tests, higher levels of fluency were associated with lower levels of lexical error gains, showing that speakers who repeated more language in the classroom activities were also making fewer errors with their lexis in the monologue speaking tests. Therefore, the conclusion reached with regard to this question is that no evidence was found to suggest that time-pressured repetition activities might lead to increases in inaccurate language use.

The final question in the study investigated whether small differences in fluency could be detected by expert raters. The results of a correlation analysis revealed that average ratings correlated significantly with three of the five fluency measures, the mean length of pause, the articulation rate, and the phonation/time ratio. This finding suggests that the small improvements of approximately 1 syllable in the mean length of run measure seen in studies such as De Jong and Perfetti (2011) and Tavakoli et al. (2016) are noticeable to human raters.

In summary, the findings of the study supported the previous research in fluency development to some degree. Over the semester, the students made small gains in decreasing their mean length of pauses, although these gains appeared to be more significant for higher proficiency learners than their lower level peers. Furthermore, these gains did not appear to negatively affect other measures of complexity and accuracy. Performance on the 3/2/1 speaking activity appeared to be related to the level of repetition that the students used, although it appeared to be limited to instances where there were larger gains in the mean length of run. Finally, although there were small but significant gains in speaking fluency, these gains were detectable to human raters.

Limitations

This study had several limitations that could have affected the results, and need therefore, to be carefully considered when interpreting the results.

First, the data were collected in classrooms where learners were able to interact with each other. Although I attempted to keep potentially distracting behavior to a minimum, in many cases there were some interactions between speakers and listeners, and such cases might have affected the data negatively, such as when listeners laughed at a comment made by a speaker, or positively, such as when a listener provided a lexical item needed by a speaker to complete their utterance. Furthermore, socialization factors might also have been an issue in student performances, as more fluent speakers might have deliberately slowed their speech rate down so as not to stand out and/or to make themselves more comprehensible to their listeners. Had the data been collected in settings where learners did not interact with each other, the findings could have been very different. Furthermore, as utterance fluency measures are affected by environmental factors, it is unclear whether any gains in cognitive fluency were made. If some measure of cognitive fluency were included in the study, the results could be very different.

Second, while the *N*-size in this study makes it one of the largest in a fluency development study, the fact remains that only 32 students of varying proficiency levels were included in this study. Because of this limitation, statistical power is an issue with the analyses conducted. The findings might have been very different had there been a larger number of students, or if they had all been of a similar level of proficiency. It should also be mentioned that some students who were absent on the days of the monologue speaking tests, and while multi-level modeling can be conducted with missing data, the results might have been different if students were present at every data collection point. Furthermore, the small sample size and range in proficiency levels meant that it was not possible to include a control group for comparison. It is unclear whether the small gains made in fluency can be attributable to the participation in the class' speaking activities, or whether students enrolled in a different kind of language course could develop in a similar fashion.

A third limitation is the assessments and the topics used for data collection. Although every effort was made to keep the test topics of comparable levels, it might have been more prudent to offer the learners a choice of topics, as some students appeared to find the topics quite difficult to speak about. Furthermore, the planning time of one minute might have been insufficient, and a different design of the speaking test might have led to a very different outcome. Similarly, while the topics were determined by the course director, some of them did not appear to be conducive to fluency development, in particular, the lesson focused on the death penalty, where students found it difficult to engage in extended discussions, and factors such as these might have affected fluency development.

Finally, due to time constraints, little data could be taken from the students themselves. It is unclear whether their personal preferences to the lesson topics and their partners could have affected their performance in the activities. Furthermore, it was only possible to account for three individual difference variables, willingness to communicate, extraversion, and L1 fluency. Had other variables, in particular, working memory, been included in the study, it might have been possible to isolate a clearer explanation of how fluency develops. Working memory could be a key factor due to the demands that speaking a foreign language with a high degree of fluency places on language processing and monitoring one's output. Other individual difference factors such as motivation or whether the students envisioned themselves using English in the future could also have explained more about their utterance fluency development.

Suggestions for Future Research

This study was exploratory in that very few researchers have longitudinally investigated fluency development in foreign language instructed settings. Based on the results, several possible directions for future research can be suggested for researchers hoping to learn more about the topic.

The first area of future research would be to replicate the study with a larger *N*-size and over a longer period of time. With a larger number of students and longer time intervals between the data collection points, it might be possible to see a clearer picture of how fluency develops. The findings of this study suggest that the proficiency level of learners might affect the degree to which fluency development activities are able to positively affect speaking skills. With a larger sample size, it might be possible to clarify this finding, and ascertain more clearly the profiles of learners who could benefit the most from fluency activities. With a larger sample size, it might also be possible to include a range of covariates such as grammatical and lexical knowledge, in addition to previous levels of exposure to the target language. Furthermore, as the length of this study was fairly short, and given that speaking fluency might take a longer time to develop for learners who do not have regular exposure to the target language, it is desirable to see if a longer time frame identifies any clearer trends. It would also be desirable to begin the study at the beginning of the academic year, as the students in this study could have followed a different trajectory of fluency development in the first semester of their studies.

One alternative to replicating the study with a larger sample size would also be to run bootstrap simulations with the current data used in this study. This statistical technique can be used to simulate replication of an analysis with a larger sample size by resampling from the current population (LaFlair, Egbert, & Plonsky, 2015). Such analyses might indicate areas of possible significant development or relationships that were obscured by the small number of students included in this study.

Including a control group in a future study might also shed some insight into the efficacy of the fluency activities used in this study. One point to investigate further is the suitability of a 3-2-1 speaking activity, especially for the lower level learners in this study who did not make a lot of progress in their fluency development. One possibility is that the task was simply too difficult for some of the students in this study. Designing some simpler fluency activities, for example based on every day experiences rather than some of the fairly abstract issues that the 3-2-1 speaking task was based around, might be more appropriate for some of the learners in this study, although the possibility remains that learners need to improve other aspects of their language skills before speaking fluency can develop. Future researchers should consider comparing similar groups of students with different fluency activities to shed some light on this issue.

Another area to investigate in more detail is that of repetition. While lexical overlap was positively correlated with some of the gains in the mean length of run, the relationship appeared to only occur when a threshold gain level had been exceeded. As the results of this study appear to suggest that repetition is only helpful in certain situations, it seems an important issue to follow up in further research. As mentioned in the previous chapter, the effect of repetition might be limited by the difficulty of the topic and/or the learner's level of engagement with it. It would be desirable to conduct follow-up interviews with students to learn more about whether these factors play a role in moderating the effect that repetition has on fluency development.

A monologue speaking test was used as a means of determining the degree to which participation in fluency development activities might transfer to other speaking tasks. As noted above, these tasks seemed quite difficult for the students, and it would be interesting to replicate the study with a wider range of assessment methods. As the task and the conditions under which it is implemented might affect the performance given by students, it would be interesting to determine whether different types of tests elicited different fluency profiles.

Finally, it would be helpful to replicate the study with a control group that did not engage in time-pressured fluency development activities. While several researchers have hypothesized that the conditions under which speaking fluency best develops (Nation & Newton, 2008; Segalowitz, 2009), it remains to be seen whether these claims are supported by empirical research. It would be desirable to see whether students who partook in fluency development activities of a different nature, for example, the fluency awareness building strategy training seen in Tavakoli et al.

(2016) developed in a similar manner to the students in this study. For speaking instruction to be most effective, it should be known what kinds of activities can help students to improve their speaking fluency and language skills.

Final Conclusions

As a teacher who has taught many classes with a speaking focus in Japanese universities, I have often heard my students say that they would like to speak English fluently. Many of them complain that they are given few opportunities to develop their speaking skills, even in classes dedicated to the skill. While this is somewhat understandable in high school contexts where many students are preparing for exams that test them on their knowledge of English grammar and vocabulary, most students at the university level have fewer constraints. In addition, many of them would like to be able to speak English to improve their job prospects after graduation. Yet despite the clear need for more knowledge about how to help students to achieve this goal, there appears to be little empirical data about the best way to go about doing so. In fact, as some researchers have mentioned, it is often assumed that by merely using speaking activities in the classroom, students will naturally improve their communication skills (Bygate, 2009). One strong reason for undertaking this research project was to try to learn more about how students can develop their speaking fluency in instructed foreign language settings, and whether some of the approaches advocated in the pedagogical literature were effective instruments for doing so.

While the results of the study do not suggest that in-class performance or individual difference variables are able to explain much about how speaking fluency develops, I do believe that the findings of the study warrant the use of these activities in such foreign language instructed settings. Importantly, students' levels of accuracy did not appear to be negatively affected by the activities, and fluency levels were stable, although some of the topics became more abstract as the semester progressed. It is also important to note that most theories of second language acquisition emphasize that language use and interaction plays a major role in the development of communicative skills. The criteria for fluency development activities, I believe, are appropriate guidelines for helping teachers to design tasks to help students interact meaningfully with each other. In many of the lessons, I saw students genuinely interested in exchanging ideas with their peers, negotiating meaning to avoid communication breakdowns, and asking questions to learn more about one another's opinions. Had the lessons been of a more traditional format, I do not feel confident that such a level of engagement would have occurred.

While many SLA researchers are engaged with the issues related to the development of complexity, accuracy, and fluency, very few of them have engaged upon longitudinal research to investigate how these variables develop over longer periods of time. While the findings of this study are modest, I hope that they add to the discussion of how learners can potentially benefit from fluency development activities. As the ability to speak fluently can be an important priority for many learners, as language teachers and researchers need to learn more about the processes involved in developing this ability. I plan to continue this line of research, and hope that others will find points of interest to investigate in their own research.

REFERENCES

Aijie, Z. (2006). A further study of the effects of the 4/3/2 technique on spoken English production. *CELEA Journal, 29*, 12-20. Retrieved from http://www.elt-china.org/teic/65/65-12.pdf

Anderson, J. R. (1992). Automaticity and the ACT* theory. *American Journal of Psychology, 105,* 165-180. doi:10.2307/1423026

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought.* Mahwah, NJ: Erlbaum.

Apple, M. T. (2011). *The big five personality traits and foreign language speaking confidence among Japanese EFL students*. Tokyo: Temple University Japan. Retrieved from http://digital.library.temple.edu/cdm/ref/collection/p245801coll10/id/127286

Arevart, S., & Nation, P. (1991). Fluency improvement in a second language. *RELC Journal, 22*, 84-94. doi:10.1177/003368829102200106

Bamford, J. (1989). Comments on Elizabeth Gatbonton and Norman Segalowitz's "creative automatization: principles for promoting fluency within a communicative framework." A reader reacts. *TESOL Quarterly, 23,* 363-366. doi:10.2307/3587360

Boers, F. (2014). A reappraisal of the 4/3/2 activity. *RELC Journal, 45*, 221-235. doi:10.1177/0033688214546964

Boersma, P. (2011). *Praat: Doing phonetics by computer* [Computer Software]. Retrieved from http://www.praat.org/

Bulte, B., & Housen, A. (2012). Defining and operationalizing L2 complexity. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy, and fluency in SLA* (pp. 21-46). Amsterdam: Benjamins.

Busch, D. (1982). Introversion-extraversion and the EFL proficiency of Japanese students. *Language Learning, 32*, 109-132. doi:10.1111/j.1467-1770.1982.tb00521.x

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* (pp. 23-48). London: Longman.

Bygate, M. (2009). Teaching and testing speaking. In M. H. Long, & C. J. Doughty (Eds.), *The handbook of language teaching* (pp. 412-440). Malden, MA: Wiley-Blackwell.

Cao, Y., & Philp, J. (2006). Interactional context and willingness to communicate: a comparison of behavior in whole class, group and dyadic interaction. *System, 34*, 480-493. doi.org/10.1016/j.system.2006.05.002

De Bot, K. (1992). A bilingual production model: Levelt's 'Speaking' model adapted. *Applied Linguistics*, *13*, 1-24. doi:10.1093/applin/13.1.1

De Jong, N. H., & Bosker, H. R. (2013). Choosing a threshold for silent pauses to measure second language fluency. In R. Eklund (Ed.), *The 6th Workshop on Disfluency in Spontaneous Speech* (DiSS) (pp. 17-20). Stockholm: Universitetsservice US-AB. Retrieved from http://pubman.mpdl.mpg.de/pubman/item/escidoc:1900232/component/escidoc:1900231/DeJong_Bosker_DiSS2013.pdf

De Jong, N. H., Groenhout, R., Schoonen, R., & Hulstijn, J. H. (2015). Second language fluency: Speaking style or proficiency? Correcting measures of second language for first language behaviour. *Applied Psycholinguistics, 32,* 223-243. doi:10.1017/S0142716413000210

De Jong, N., & Perfetti, C. A. (2011). Fluency training in the ESL classroom: An experimental study of fluency development and proceduralization. *Language Learning*, *61*, 533-568. doi:10.1111/j.1467-9922.2010.00620.x

De Jong, N. H., & Steinel, M. P., Florijin, A., Schoonen, R., & Hulstijn, J. H. (2012a). Linguistic skills and speaking fluency in a second language. *Applied Psycholinguistics, 34,* 893-916. doi:10.1017/S0142716412000069

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

De Jong, N. H., & Steinel, M. P., Florijin, A., Schoonen, R., & Hulstijn, J. H. (2012c). Facets of speaking proficiency. *Studies in Second Language Acquisition, 34,* 4-34. doi:10.1017/S0272263111000489

De Jong, N. H., & Vercellotti, M. L. (2016). Similar prompts may not be similar in the performance they elicit: Examining fluency, complexity, accuracy, and lexis in narratives from five picture prompts. *Language Teaching Research, 20,* 387-404. doi:10.1177/1362168815606161

Delany, T. (2012). Quality and quantity of oral participation and English proficiency gains. *Language Teaching Research*, *16*, 467-482. doi:10.1177 /1362168812455586

DeKeyser, R. (1997). Beyond explicit rule learning: Automatizing second language morphosyntax. *Studies in Second Language Acquisition, 19*, 195-221.

DeKeyser, R. (2007a). Introduction: Situating the concept of practice. In R. DeKeyser (Ed.) *Practice in a second language* (pp. 1-18). Cambridge: Cambridge University Press.

DeKeyser, R. (2007b). Skill acquisition theory. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition* (pp. 97-114). Mahwah, NJ: Erlbaum.

DeKeyser, R. (2010). Monitoring processes in Spanish as a second language during a study abroad program. *Foreign Language Annals*, *43*, 80-92. doi:10.1111 /j.1944-9720.2010.01061.x

Derwing, T. M., Munro, M. J., & Thomson, R. I. (2007). A longitudinal study of ESL learners' fluency and comprehensibility development. *Applied Linguistics, 29*, 359-380. doi:10.1093/applin/amm041

Derwing, T. M., Munro, M. J., Thomson, R. I., & Rossiter, M. J. (2009). The relationship between L1 fluency and L2 fluency development. *Studies in Second Language Acquisition, 31,* 533-557. doi:10.1017/S0272263109990015

Derwing, T. M., & Munro, M. J. (2013). The development of L2 oral language skills in two L2 groups: A 7 year study. *Language Learning, 63*, 163-185. doi:10.1111/lang.12000

Dewaele, J., & Furnham, A. (1999). Extraversion: The unloved variable in applied linguistic research. *Language Learning, 49*, 509-544. doi:10.1111/0023-8333.00098

Dewaele, J., & Furnham, A., (2000). Personality and speech production: a pilot study of second language learners. *Personality and Individual Differences, 28*, 355-365. doi:http://dx.doi.org.libproxy.temple.edu/10.1016/S0191-8869(99)00106-3

Doe, T. (2013). Willingness to communicate and group cohesion. In N. Sonda & A. Krause (Eds.), *JALT 2013 Conference Proceedings.* Tokyo: JALT. Retrieved from http://jalt-publications.org/proceedings/articles/4001-selected-paper-willingness-communicate-and-group-cohesion

Dörnyei, Z. (2009). *The psychology of second language acquisition.* Oxford: Oxford University Press.

Dörnyei, Z., & Kormos, J. (2000). The role of individual and social variables in oral task performance. *Language Teaching Research, 4*, 275-300. doi:10.1177 /136216880000400305

Dörnyei, Z., & Ryan, S. (2015). *The psychology of the language learner revisited.* New York, NY: Routledge.

Du, H. (2013). The development of Chinese fluency during study abroad in China. *Modern Language Journal, 97*, 131-143. doi:10.1111/j.1540-4781.2013 .01434.x

Ellis, R. (2003). *Task-based language learning and teaching*. Oxford: Oxford University Press.

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

Ellis, R. (2008). *The study of second language acquisition* (2nd ed.). Oxford: Oxford University Press.

Ellis, R. (2009). The differential effects of three types of task planning on the fluency, complexity, and accuracy in L2 oral production. *Applied Linguistics, 30*, 474-509.

Ellis, R., & Barkhuizen, G. (2005). *Analyzing learner language*. Oxford: Oxford University Press.

Elwood, J. (2011). *Enriching structural models of L2 willingness to communicate: the role of personality, ego, permeability, and perceived distance.* Tokyo: Temple University Japan. Retrieved from http://digital.library.temple.edu/cdm/ref/collection/p245801coll10/id/134212

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). London: Sage.

Fisher, W. P. Jr. (2007). Rating scale instrument quality criteria. *Rasch Measurement Transactions, 21*, 1095.
Retrieved from: https://www.rasch.org/rmt/rmt211m.htm

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

Freed, B. F., Segalowitz, N., & Dewey, D. (2004). Context of learning and second language fluency in French: Comparing regular classroom, study abroad, and intensive domestic immersion programs. *Studies in Second Language Acquisition, 26*, 275-301. doi:10.1017/S0272263104262064

Fukuta, J. (2016). Effects of task repetition on learners' attention orientation in L2 oral production. *Language Teaching Research, 30*, 321-340. doi:10.1177/1362168815570142

Gatbonton, E., & Segalowitz, N. (1988). Principles for promoting fluency within a communicative framework. *TESOL Quarterly, 22*, 473-492. doi:10.2307/3587290

Gatbonton, E., & Segalowitz, N. (1989). Comments on Elizabeth Gatbonton and Norman Segalowitz's "creative automatization: principles for promoting fluency within a communicative framework." The authors respond. *TESOL Quarterly, 23,* 366-373. doi:10.2307/3587361

Gatbonton, E., & Segalowitz, N. (2005). Rethinking communicative language teaching: A focus on access to fluency. *The Canadian Modern Language Journal*, *61*, 325-353. doi:10.3138/cmlr.61.3.325

Gibson, J. J. (1986). *The ecological approach to visual perception.* Hillsdale, NJ: Erlbaum.

Goldman Eisler, F. (1968). *Psycholinguistics: Experiments in spontaneous speech*. London: Academic Press.

Goldstein, H. (2011). *Multilevel statistical models* (4th ed.). West Sussex: Wiley.

Guillot, M. (1999). *Fluency and its teaching.* Clevedon: Multilingual Matters.

Hieke, A. E. (1985). A componential approach to oral fluency evaluation. *Modern Language Journal, 69*, 135-142. doi:10.1111/j.1540-4781.1985.tb01930.x

Housen, A., & Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics*, *30*, 461-473. doi:10.1093/applin/amp048

Housen, A., Kuiken, F., & Vedder, I. (2012). Complexity, accuracy, and fluency: Definitions, measurement and research. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy, and fluency in SLA* (pp. 1-20). Amsterdam: Benjamins.

Jarvis, S. (2013). Capturing the diversity in lexical diversity. In L. Ortega, A. Cumming, & N. C. Ellis (Eds.), *Agendas for language learning research* (pp. 87-106). Malden, MA: Wiley-Blackwell.

Kahng, J. (2014). Exploring utterance and cognitive fluency of L1 and L2 English speakers: Temporal measures and stimulated recall. *Language Learning, 64*, 809-854. doi:10.1111/lang.12084

Kang, S. (2005). Dynamic emergence of situational willingness to communicate in a second language. *System, 33*, 399-407.

Kormos, J. (2006). *Speech production and second language acquisition.* New York, NY: Routledge. doi.org/10.1016/j.system.2004.10.004

Kormos, J., & Denes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System, 32*, 145-164. Retrieved from http: //dx.doi.org.libproxy.temple.edu/10.1016/j.system.2004.01.001

LaFlair, T., Egbert, J., & Plonsky, L. (2015). A practical guide to bootstrapping descriptive statistics, correlations, *t* tests, and ANOVAs. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research.* New York, NY: Routledge.

Lambert, C., Kormos, J., & Minn, D. (2016). Task repetition and second language speech processing. *Studies in Second Language Acquisition*. doi:10.1017/S0272263116000085

Larsen-Freeman, D. (2006). The emergence of fluency, accuracy and complexity in the oral and written production of five Chinese learners of English. *Applied Linguistics, 27*, 590-619. doi:10.1093/applin/aml029

Larsen-Freeman, D. (2009). Adjusting expectations: The study of complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics, 30*, 579-589. doi:10.1093/applin/amp043

Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning, 40*, 387-417. doi:10.1111/j.1467-1770.1990.tb00669.x

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

Lightbown, P. (2007). Transfer appropriate processing as a model for classroom second language acquisition. In Z. Han (Ed.), *Understanding second language process* (pp. 27-44). Clevedon: Multilingual Matters.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement, 3*, 85-106.

Linacre, J. M. (2015). *WINSTEPS Rasch measurement computer program* (Version 3.91.0). Chicago: Winsteps.com.

Logan, G. D. (2002). Toward an instance theory of automatization. *Psychological Review, 95,* 492-527. doi:10.1037/0033-295X.95.4.492

Lowie, W., & Verspoor, M. (2011). The dynamics of multilingualism: Levelt's speaking model revisited. In M. S. Schmid & W. Lowie (Eds.), *Modeling bilingualism: From structure to chaos* (pp. 267-288). Amsterdam: Benjamins.

Maurice, K. (1983). The fluency workshop. *TESOL Newsletter, 17,* 429.

MacIntyre, P. D., Baker, S., Clement, R., and Conrod, S. (2001). Willingness to communicate, social support, and language-learning orientations of immersion students. *Studies in Second Language Acquisition, 23*, 369-388. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.542.9753&rep=rep1&type=pdf

MacWhinney, B. (2000). *The CHILDES project: tools for analyzing talk* (3rd ed.). Mahwah, NJ: Erlbaum.

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods, 42*, 381-392. doi:10.3758/BRM.42.2.381

McLaughlin, B., Rossman, T., & McLeod, B. (1983). Second language learning: An information processing perspective. *Language Learning, 33*, 135-157. doi:10.1111/j.1467-1770.1983.tb00532.x

McNamara, T. F. (1996). *Measuring second language performance*. London: Longman.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). Automated evaluation of text and discourse with Coh-Metrix. Cambridge, MA: Cambridge University Press.

Meiden, W. (1958). The development of fluency in foreign language courses. *Modern Language Journal, 42*, 82-86.

Mojavezi, A. (2013). The relationship between task repetition and language proficiency. *Applied Research on English Language, 3*, 29-40. Retrieved from http://www.sid.ir/en/VEWSSID/J_pdf/5064020140503.pdf

Mora, J. C., & Valls-Ferrer, M. (2012). Oral fluency, accuracy, and complexity in formal instruction and study abroad learning contexts. *TESOL Quarterly, 46*, 610-641. doi:10.1002/tesq.34

Munezane, Y. (2014). *A structural equation model and intervention study of individual differences, willingness to communicate, and L2 use in an EFL classroom.* Tokyo: Temple University Japan. Retrieved from http://cdm16002.contentdm.oclc.org/cdm/compoundobject/collection/p245801coll10/id/290534/rec/

Murphy, J. (2003). Task-based learning: The interaction between tasks and learners. *ELT Journal, 57*, 352-360. doi:10.1093/elt/57.4.352

Naiman, N., Frolich, M., & Stern, H. H. (1996). *The good language learner.* Clevedon: Multilingual Matters.

Nation, I. S. P. (1989). Improving speaking fluency. *System, 17*, 377-384. Retrieved from http://dx.doi.org.libproxy.temple.edu/10.1016/0346-251X(89)90010-9

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Nation, I. S. P., & Newton, J. (2009). *Teaching ESL/EFL listening and speaking*. New York, NY: Routledge.

Nitta, R., & Nakatsuhara, F. (2014). A multifaceted approach to investigating pre-task planning effects on paired oral test performance. *Language Testing, 31*(2), 147-175. doi.org/10.1177/0265532213514401

O'Brien, I., Seglowitz, N., Freed, B., & Collentine, J. (2007). Phonological memory predicts second language oral fluency gains in adults. *Studies in Second Language Acquisition, 29,* 557-582. doi:10.1017/S027226310707043X

O'Connell, A. A., & McCoach, D. B. (2008). *Multilevel modeling of educational data.* Charlotte, NC: Information Age.

Oz, H. (2014). Big five personality traits and willingness to communicate among foreign language learners in Turkey. *Social Behavior and Personality, 42*, 1473-1482. doi.org/10.2224/sbp.2014.42.9.1473

Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. C. Richards & R. W. Schmidt (Eds.), *Language and communication* (pp. 191-225). London: Longman.

Pickering, M. J., & Branigan, H. P. (1999). Syntactic priming in language production. *Trends in Cognitive Science, 3*, 136-141. Retrieved from http://dx .doi.org.libproxy.temple.edu/10.1016/S1364-6613(99)01293-0

Pickering, M. J., & Ferreira, V. S., (2008). Structural priming: A critical review. *Psychological Bulletin, 134*, 427-459. doi:10.1037/0033-2909.134.3.427

Pulvermuller, F. (2002). *The neuroscience of language* Cambridge: Cambridge University Press.

Rasch, G. (1960). *Probabilistic models for some intelligence and achievement tests*. Copenhagen: Danish Institute for Educational Research.

Raudenbush, S. W., Bryk, A. S, & Congdon, R. (2013). *HLM 7.01 for Windows* [Computer software]. Skokie, IL: Scientific Software.

Roberts, B., & Kirsner, K. (2000). Temporal cycles in speech production. *Language and Cognitive Processes, 15*, 129-157. doi:10.1080/016909600386075

Riazantseva, A. (2001). Second language proficiency and pausing: A study of Russian speakers of English. *Studies in Second Language Acquisition, 23,* 497-526.

Richards, J. C. (1990). *The language teaching matrix.* Cambridge: Cambridge University Press

Richards, J. C., & Schmidt, R. (2002). *Dictionary of language teaching and applied linguistics* (3rd ed.). London: Longman.

Richards, J., & Rodgers, T. (2001). *Approaches and methods in language teaching* (2nd ed.). Cambridge: Cambridge University Press.

Robinson, P. (2001). Task complexity, cognitive resources, and syllabus design. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 287-318). Cambridge: Cambridge University Press.

Rossiter, M. J., Derwing, T. M., Manimtim, L. G., & Thomson, R. I. (2010). Oral fluency: The neglected component in the communicative language classroom. *Canadian Modern Language Journal, 66*, 583-606. doi:10.3138/cmlr.66.4.583

Schmidt, R. (1992). Psychological mechanisms underlying second language fluency. *Studies in Second Language Acquisition*, *14*, 357-385. Retrieved from http: //nflrc.hawaii.edu/PDFs/SCHMIDT%20Psychological%20mechanisms%20underlying%20second%20language%20fluency.pdf

Segalowitz, N. (2003). Automaticity and second languages. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 382-408). Malden, MA: Blackwell.

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

Segalowitz, N., & Freed, B. (2004). Context, contact, and cognition in oral fluency acquisition. Learning Spanish in at home or study abroad contexts. *Studies in Second Language Acquisition, 26,* 173-200. doi:10.1017/S0272263104262027

Segalowitz, N., Gatbonton, E., & Trofimovich, P. (2009). Links between ethnolinguistic affiliation, self-related motivation and second language fluency: Are they mediated by psycholinguistic variables? In Z. Dörnyei, & E. Ushioda (Eds.), *Motivation, language identity, and the L2 self* (pp. 172-192). Bristol: Multilingual Matters.

Segalowitz, N., & Segalowitz, S. J. (1993). Skilled performance, practice, and the differentiation of speed-up from automatization effects: Evidence from second language word recognition. *Applied Psycholinguistics, 14,* 369-385. doi:10. 1017/S0142716400010845

Shehadeh, A. (2012). Introduction: broadening the perspective of task-based language teaching scholarship: The contribution of research in foreign language contexts. In A. Shehadeh & C. A. Coombes (Eds.), *Task-based language teaching in foreign language contexts* (pp. 1-20). Amsterdam: Benjamins.

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

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, and lexis. *Applied Linguistics, 30*, 510-532. doi:10.1093/applin /amp047

Skehan, P., & Foster, P. (2005). Strategic and on-line planning: The influence of surprise information and task time on second language performance. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 193-216). Amsterdam: Benjamins.

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

Smith, M., & Wheeldon, L., (2001). Syntactic priming in spoken sentence production: An online study. *Cognition, 78,* 123-164. Retrieved from http://dx.doi.org. libproxy.temple.edu/10.1016/S0010-0277(00)00110-4

Tavakoli, P. (2011). Pausing patterns: differences between L2 learners and native speakers. *ELT Journal, 65*, 71-79. doi:10.1093/elt/ccq020

Tavakoli, P., Campbell, C., & McCormack, J. (2016). Development of speech fluency over a short period of time: Effects of pedagogic intervention. *TESOL Quarterly, 50*, 447-471. doi:10.1002/tesq.244

Tavakoli, P., & Skehan, P. (2005). Strategic planning, task structure, and performance testing. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 239-273). Amsterdam: Benjamins.

Tavakoli, P., & Foster, P. (2008). Task design and second language performance: The effect of narrative type on learner output. *Language Learning*, *58*, 439-473. doi:10.1111/j.1467-9922.2008.00446.x

Thai, C. & Boers, F. (2016). Repeating a monologue under increasing time pressure: Effects on fluency, complexity, and accuracy. *TESOL Quarterly*, *50*, 369-393. doi:10.1002/tesq.232

Tonkyn, A. (2012). Measuring and perceiving changes in oral complexity, accuracy, and fluency: examining instructed learners' short-term gains. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy, and fluency in SLA* (pp. 221-246). Amsterdam: Benjamins.

Tomasello, M. (2003). *Constructing a language*. Cambridge, MA: Harvard University Press.

Towell, R. (2002). Relative degrees of fluency: a comparative case study of advanced learners of French. *IRAL, 40,* 117-150. doi:10.1515/iral.2002.005

Towell, R., Hawkins, & R., Bazergui, N. (1996). The development of fluency in advanced learners of French. *Applied Linguistics, 17*, 84-119. doi:10.1093 /applin/17.1.84

Van Daele, S., Housen, A., Pierrad, M., & Debruyn, L. (2006). The effect of extraversion on oral L2 proficiency. *EUROSLA Yearbook, 6*, 213-236. Retrieved from http://dx.doi.org/10.1075/eurosla.6.13dae

Van Lier, L. (2000). From input to affordances: Socio-interactive learning from an ecological perspective. In J. Lantolf (Ed.), *Sociocultural theory and second language learning* (pp. 245-259). Oxford: Oxford University Press.

Verspoor, M., Lowie, van Geert, P., van Dijk, & Schmid, M. S. (2011). How to sections. In M. Verspoor, K. de Bot, & W. Lowie (Eds.), *A dynamic approach to second language development.* Amsterdam: Benjamins.

Weaver, C. (2010). *Japanese university students' willingness to use English with different interlocutors*. Tokyo: Temple University Japan. Retrieved from http://digital.library.temple.edu/cdm/ref/collection/p245801coll10/id/65448

Wood, D. (2001). In search of fluency: What is it and how can we teach it? *Canadian Modern Language Journal, 57*, 573-589. doi:10.3138/cmlr.57.4.573

Wood, D. (2009). Effects of focused instruction of formulaic sequences on fluent expression in second language narratives: a case study. *Canadian Journal of Applied Linguistics, 12*, 39-56. Retrieved from http://search.proquest.com /docview/216969938?accountid=14270

Wood, D. (2010). *Formulaic language and second language speech fluency.* London: Continuum.

Wray, A. (2002). *Formulaic language and the lexicon.* Cambridge: Cambridge University Press.

Wray, A. (2008). *Formulaic language: Pushing the boundaries.* Oxford: Oxford University Press.

APPENDICES

APPENDIX A
LIST OF FORUMULAIC PHRASES DIRECTLY TAUGHT IN THE DISCUSSION COURSE

1. **First Semester, Level 1**

Lesson 2: Opinions and Reasons	
Asking for Opinions and Reasons	Giving Opinions and Reasons
What’s your opinion?	Personally speaking, I think…
What does everyone (else) think?	I’m not sure, but I think…
How come?	The main reason is…
Can you tell me why?	One reason is…

Lesson 3: Organizing Discussions	
Joining a Discussion	Connecting Ideas
Can I start?	As {you / name} said,…
Can I make a comment?	What do you think of {my / name’s} idea?
Can I ask a question?	Does anyone agree with {my / name’s} idea?

Lesson 6: Examples and Experiences	
Asking for Examples and Experiences	Giving Examples and Experiences
Can you give me an example?	One example is…
For example / instance?	For example / instance,…
Have you ever… ?	When I was… / Last year, I…

Lesson 7: Changing Topic	
Checking If Everyone’s Finished	Changing Topic
Does anyone want to comment?	What shall we discuss first / next?
Does anyone want to add something?	Why don’t we discuss {topic}?
Does anyone want to ask a question?	

Lesson 10: Alternative Ideas	
Asking for Alternative Ideas	Giving Alternative Ideas
Does anyone have an alternative idea?	One alternative could be...
Can anyone think of a different idea?	Another alternative could be…

Lesson 11: Different Viewpoints	
Asking for Different Viewpoints	Giving Different Viewpoints
How about for {X}?	For {X},…
How about from {X}’s point of view?	From {X}’s point of view,…
How about from {X}’s perspective?	From {X}’s perspective,…

2. **First Semester, Levels 2-3**

Lesson 2: Opinions	
Asking for Opinions	Giving Opinions
What's your opinion?	In my opinion,…
What do you think?	Personally speaking, I think…
What does everyone think?	I'm not sure, but I think…

Lesson 3: Reasons	
Asking for Reasons	Giving Reasons
Why {do you think so}?	It's mainly because…
How come?	One reason is…
Can you tell me why?	Another reason is…

Lesson 6: Examples	
Asking for Examples	Giving Examples
For example / instance?	For example / instance,…
Can you give me an example?	One example is…
Such as?	Another example is…

Lesson 7: Joining a Discussion	
Joining a Discussion	Asking Others to Join a Discussion
Can I start?	
Can I make a comment?	Does anyone want to comment?
Can I add something?	Does anyone want to add something?
Can I ask a question?	

Lesson 10: Connecting Ideas	
Asking Others to Connect	Connecting to Others' Opinions
What do you think of {my / name's} idea?	As {you / name} said,…
Does anyone agree with {my / name's} idea?	Like {you / name} said,…

Lesson 11: Experiences	
Asking for Experiences	Talking about Experiences
Have you ever… ?	When <u>I was in high school</u>, I…
When {did you}… ?	<u>Yesterday</u> / <u>Last week</u> / <u>Three years ago</u>, I…
How often {do you}… ?	<u>Sometimes</u> / <u>Often</u> / <u>Every day</u>, I…

3. First Semester, Level 4

Opinions	
Asking for Opinions	Giving Opinions
What do you think?	I think…
What's your opinion?	In my opinion,…
What does everyone think?	Personally speaking, I think…

Reasons	
Asking for Reasons	Giving Reasons
Why?	Because…
How come?	It's mainly because…
Can you tell me why?	One reason is…

Examples	
Asking for Examples	Giving Examples
For example?	For example,…
For instance?	For instance,…
Can you give me an example?	One example is…

Joining a Discussion	
Joining a Discussion	Asking Others to Join a Discussion
Can I start?	
Can I make a comment?	Does anyone want to comment?
Can I ask a question?	

Connecting Ideas	
Asking Others to Connect	Connecting to Others' Opinions
What do you think of {my / name's} idea?	As {you / name} said,…
	Like {you / name} said,…

Experiences	
Asking for Experiences	Talking about Experiences
Have you ever… ?	When I was in high school, I…
When {did you}… ?	Yesterday / Last week / Three years ago, I…
How often {do you}… ?	Sometimes / Often / Everyday, I…

4. **Second Semester, Level 1**

Lesson 2: Surveying and Summarizing	
Surveying	Summarizing
Who thinks (that)… ?	So, we all think…
Does anyone else think (that)… ?	So, most / some of us think…
	So, it seems we can't agree.

Lesson 3: Possibilities	
Asking about Possibilities	Talking about Possibilities
If… ?	If…

Lesson 6: Avoiding Ambiguity	
Listeners	Speakers
What do you mean by {X}?	By {X}, I mean…
How would you define {X}?	I would define {X} as…

Lesson 7: Reporting Information	
Asking Others to Report Information	Reporting Information
Where did you hear that?	I heard / read / saw (that)…
How do you know about that?	{Name} said (that)…

Lesson 10: Majority and Minority Views	
Listeners	Speakers
Do you think most people would agree?	Like most people, I think…
Do you think some people would disagree?	Unlike most people, I think…
	Some people think…

Lesson 11: Balancing Opinions	
Helping Others to Balance	Balancing Your Opinion
What are the advantages (of…)?	One / Another advantage (of…) is…
What are the disadvantages (of…)?	One / Another disadvantage (of…) is…

5. Second Semester, Levels 2-3

Lesson 2: Changing Topic	
Checking If Everyone's Finished	Changing Topic
Does anyone want to comment?	What shall we discuss first / next?
Does anyone want to ask a question?	Why don't we discuss {topic}?

Lesson 3: Possibilities	
Asking about Possibilities	Talking about Possibilities
If…?	If…

Lesson 6: Paraphrasing	
Paraphrasing Others	Paraphrasing Yourself
Do you mean (that)…?	I mean (that)…
So what you're saying is (that)…?	What I'm saying is (that)…
So, in other words… ?	In other words…

Lesson 7: Reporting Information	
Asking Others to Report Information	Reporting Information
Where did you hear that?	I heard / read / saw (that)…
How do you know about that?	{Name} said (that)…

Lesson 10: Different Viewpoints	
Asking for Different Viewpoints	Giving Different Viewpoints
How about for {X}?	For {X},…
How about from {X}'s point of view?	From {X}'s point of view,…

Lesson 11: Balancing Opinions	
Asking Others to Balance Opinions	Balancing Your Opinions
What are the advantages (of…) ?	One / Another advantage is (that)…
What are the disadvantages (of…) ?	One / Another disadvantage is (that)…

6. Second Semester, Level 4

Lesson 2: Changing Topic	
Checking If Everyone's Finished	Changing Topic
Does anyone want to comment?	What shall we discuss first / next?
Does anyone want to ask a question?	Why don't we discuss {topic}?

Lesson 3: Possibilities	
Asking about Possibilities	Talking about Possibilities
If…?	If…

Lesson 6: Paraphrasing	
Paraphrasing Others	Paraphrasing Yourself
Do you mean (that)… ?	I mean (that)…
So, in other words… ?	In other words…

Lesson 7: Reporting Information	
Asking Others to Report Information	Reporting Information
Where did you hear that?	I heard / read / saw (that)…
How do you know about that?	{Name} said (that)…

Lesson 10: Different Viewpoints	
Asking for Different Viewpoints	Giving Different Viewpoints
How about for {X}?	For {X},…
How about from {X}'s point of view?	From {X}'s point of view,…

Lesson 11: Balancing Opinions	
Asking Others to Balance Opinions	Balancing Your Opinions
What are the advantages (of…) ?	One / Another advantage is (that)…
What are the disadvantages (of…) ?	One / Another disadvantage is (that)…

APPENDIX B
SAMPLE QUIZ (LESSON 3, FALL SEMESTER)

1. If English was an official language in Japan, what does the article say will be provided in Japanese and English?

a) Restaurant service
b) Driver's licenses
c) Medical care
d) Speeches by politicians

2. Which two organizations use English as an official language?

a) ACT UP and Africare
b) Earth Justice and Eurostep
c) The IMF and International Pen Friends
d) The United Nations and UNICEF

3. About how long does it take children to master Japanese?

a) About seven years
b) About eight years
c) About nine years
d) About ten years

4. From 2011, which grades in elementary school must study English?

a) fifth and sixth grades
b) sixth grade only
c) fifth grade only
d) all grades

5. The Japanese governments wants elementary school students to have…

a) basic English communication skills.
b) knowledge of English grammar.
c) good listening skills.
d) clear English pronunciation.

6. Who are mentioned as being glad that English is being taught in elementary schools?

a) Students
b) English teachers
c) Politicians
d) Parents

7. How does the 30 year-old businessman feel about Rakuten's English-only policy?

a) Worried
b) Proud
c) Angry
d) Excited

8. What did one Japanese student say about using English in Japanese companies?

a) More companies will use English.
b) Rakuten will never use 100% English.
c) Companies will need to hire English teachers.
d) Having English speakers will not make a company successful.

APPENDIX C
SAMPLE FORM-FOCUSED PRACTICE OF COMMUNICATIVE FUNCTIONS (CHANGING TOPIC: LESSON 3, FALL SEMESTER)

Controlled Practice

Fill in each space below with one of today's function phrases.

Ryo: Personally speaking, I think English classes should be taught only by native teachers because it's helpful to listen to natural pronunciation. ______________________________?

Aki: Yes. I see your point, but I'm not sure I agree. Some of my teachers are not native English speakers, but they speak English really well. I want to be able to speak like them! ______________________________?

Ryo: No.

Jun: I'm okay.

Ryo: Okay. ______________________________?

Jun: ______________________________ other foreign language classes?

All: Sure.

Jun: Okay, do you think it's a good idea that we study one other foreign language in addition to English?

Semi-controlled practice

Discuss the following questions. Use today's phrases to check if everyone's finished and change topic.

- *(Topic: "Getting a Good Job")* Will learning English help you get a good job in the future?
- *(Topic: "Learning English")* Do you enjoy learning English?
- *(Topic: "Travel")* Do you want to travel to an English speaking country?
- *(Topic: "Helping Others")* Would you help foreign tourists lost in Tokyo?
- *(Topic: "English Skills")* Is it more important to develop speaking or reading skills in English?

APPENDIX D
REGULAR LESSON DISCUSSION PREPARATION TASKS

Session 1 (Lesson 3): When to Study Abroad
Preparation
1. *Below are four ideas about studying abroad. For each opinion, decide if you agree or disagree.*

Ryo: I think the best time to study abroad is in **high school**. If we study abroad in high school, we can stay with a host family and learn about foreign culture deeply.

(agree / disagree)

Aki: In my opinion, the best time to study abroad is at **university**. If we study abroad at university, we can get a good job when we graduate because companies want people who can speak foreign languages.

(agree / disagree)

Jun: Personally speaking, I think the best time to study abroad is **after working for a couple of years**. If we work and save money first, we don't have to ask our parents for money.

(agree / disagree)

Eri: I think people should **never study abroad**. If we want to communicate with foreign people and learn foreign cultures, it's easier to just use the Internet.

(agree / disagree)

Session 2 (Lesson 4): Being Fashionable
Preparation
1. *Below are four opinions about fashion. For each opinion, decide if you agree or disagree.*

Ryo: I think **comfortable clothes** are best! I often wear jeans and t-shirts because I can feel relaxed.

(agree / disagree)

Eri: In my opinion, it's fun to wear **fashionable clothes**. I enjoy reading fashion magazines and talking about clothes with my friends.

(*agree / disagree)*

Jun: I don't follow fashion because I want to have my own **individual style**. I sing in a band, so I wear black clothes and dye my hair. Some people don't like my style, but I don't care.

(agree / disagree)

Aki: I feel a lot of **pressure to be fashionable** at university. It takes me a long time to choose what to wear every morning, and I sometimes worry that I don't look fashionable.

(agree / disagree)

Session 3 (Lesson 6): Information and the Internet
Preparation

1. *Below are four opinions about the Internet. For each opinion, decide if you agree or disagree.*

Eri: **Smart phones** are great because we can use them to get any information at any time. We can surf the Internet and check the news on the train, in a cafe, or anywhere.

(agree / disagree)

Jun: I think **books, magazines, and newspapers** are better than the Internet. The information is deeper and more accurate.

(agree / disagree)

Aki: When I do homework, I can find useful information on websites like **Wikipedia**. The Internet has lots of information to help me think about my ideas.

(agree / disagree)

Ryo: Many of my friends watch **YouTube** for hours every day! I think people spend too much time on the Internet.

(agree / disagree)

Session 4 (Lesson 7): How to Become Famous
Preparation

1. *Below are four opinions about the becoming famous. For each opinion, decide if you agree or disagree.*

Eri: I think that being hard working is the most important quality if you want to become famous. Most celebrities work very hard and can't take many holidays.

(agree / disagree)

Jun: In my opinion, it's important to be good looking and fashionable if you want to be famous. All the popular celebrities are cute or handsome.

(agree / disagree)

Aki: I think that you need to be very lucky if you want to be famous. There are many talented people who want to become famous, so you need luck to be successful.

(agree / disagree)

Ryo: I think you need to be creative and intelligent if you want to become famous. Famous celebrities always say something funny or clever when they are interviewed.

(agree / disagree)

Session 5 (Lesson 8): Japanese Traditional Culture
Preparation

1. *Below are four opinions about Japanese traditional culture. For each opinion, decide if you agree or disagree.*

Eri: In my opinion, it's important for Japanese people to learn about Japanese traditional culture. For instance, if we study calligraphy (*shodo*) or tea ceremony, we can understand our culture more deeply.

(agree / disagree)

Jun: I think that Japanese traditional culture is dying. Only old people are interested in ikebana, *shogi*, and other traditional culture. Young people are more interested in pop culture.

(agree / disagree)

Aki: Japanese traditional culture is important for the economy. Many foreign tourists are interested in Japanese traditional culture, so they pay a lot of money to visit places like Kyoto and Nara.

(agree / disagree)

Ryo: Japanese traditional culture is too difficult. It's really hard to understand kabuki or old Japanese literature, like the *Kojiki* and *Genji Monogatari.*

(agree / disagree)

Session 6 (Lesson 10): Public Rules
Preparation

1. *Below are four rules about what people can do in public places. Do you agree or disagree with each rule?*

Australia: No Smoking in Public
Rule - People can't smoke at beaches, parks, sports events, or restaurants
Punishment - A fine of ¥20,000 for smokers

(agree / disagree)

India: No Kissing in Public
Rule - People should not kiss in public
Punishment - Couples pay fines and can be sent to prison for three months

(agree / disagree)

Taiwan: No Noisy Neighbors
Rule - People should not make too much noise in their neighborhood
Punishment - A fine of ¥50,000 for people who make too much noise
(agree / disagree)

Washington, D.C., USA: No Eating and Drinking on Trains
Rule - People can't eat or drink on trains or at stations
Punishment - A fine of ¥10,000 and police can arrest people who eat or drink
(agree / disagree)

Session 7 (Lesson 11): The Death Penalty
Preparation

1. *Below are four opinions about the death penalty. For each opinion, decide if you agree or disagree.*

Jun: I think the death penalty should be used for people who do really bad things, like killing children. If murderers are punished by the death penalty, then ordinary people will feel safer.
(agree / disagree)

Ryo: All criminals deserve a second chance. Even if people commit murder, they can change in the future. The death penalty is wrong.
(agree / disagree)

Eri: The death penalty is a good punishment because it prevents crime. Most people will be too scared to commit murder if they know about the death penalty.
(agree / disagree)

Aki: Sometimes the law makes mistakes, and innocent people are sentenced to death. We should never use the death penalty because serious mistakes can be made.
(agree / disagree)

Session 8 (Lesson 12): Gender Stereotypes
Preparation

1. *Below are four opinions about men and women. For each opinion, decide if you agree or disagree.*

- Men are not good at listening. (*agree / disagree*)
- Men are better leaders than women. (*agree / disagree*)
- Women are bad at math and science. (*agree / disagree*)
- Women are kinder than men. (*agree / disagree*)

APPENDIX E
REGULAR LESSON DISCUSSION QUESTIONS

Session 1 (Lesson 3): When to Study Abroad

1. What are the advantages and disadvantages of studying abroad:
   - in high school?
   - in university?
   - after working for a couple of years?
2. Should all students study abroad?

Session 2 (Lesson 4): Being Fashionable

1. Do you think it's important to wear:
   - comfortable clothes?
   - fashionable clothes?
   - your own individual style?
2. Do university students have too much pressure to be fashionable?

Session 3 (Lesson 6): Information and the Internet

1. Do you think the Internet is a good way to get information?
2. What is the best way to get information: from TV, the Internet, books, or newspapers?

Session 4 (Lesson 7): How to Become Famous

1. Which qualities do you think are most important to be a famous:
   - sports star?
   - TV presenter?
   - singer / musician?
2. To become famous, do you think all celebrities have to be good-looking?

Session 5 (Lesson 8): Japanese Traditional Culture

1. Which would you prefer to introduce to a foreign friend: Japanese pop culture or Japanese traditional culture?
2. Do you think traditional culture is more important than pop culture?

Session 6 (Lesson 10): Public Rules

1. What are the advantages and disadvantages of public rules?
2. Which public rules should be introduced to Tokyo?

Session 7 (Lesson 11): The Death Penalty

1. What are the advantages and disadvantages of the death penalty?
2. Do you think the death penalty is a good way to punish murderers?

Session 8 (Lesson 12): Gender Stereotypes

1. What stereotypes of men and women do you know?
2. Do you think these stereotypes are true?

APPENDIX F
3/2/1 TASK FLUENCY QUESTIONS

1. What opinions and ideas did you share in today's class?
2. What opinions and ideas did you hear in today's class?

APPENDIX G
TWO-MINUTE MONOLOGUE TEST QUESTIONS (SHORT-TERM FLUENCY MEASURE)

1. Do you prefer traveling in Japan or abroad? Why?
2. Do you prefer the city or the country? Why?
3. Do you prefer watching Youtube or watching movies? Why?
4. Do you prefer TV or computer games? Why?
5. Do you prefer playing or watching sport? Why?
6. Do you prefer playing or listening to music? Why?
7. Do you prefer summer or winter? Why?
8. Do you prefer birthdays or New Year? Why?
9. Do you prefer eating at home or eating in a restaurant? Why?
10. Do you prefer spending time alone or with friends? Why?

APPENDIX H
TWO-MINUTE L1 FLUENCY QUESTIONS (JAPANESE FLUENCY MEASURE)

1. お金を貯金することか使う方、どちらを好みますか？それはなぜですか？
[Do you prefer to save or spend money? Why?]

2.暑い天気か寒い天気、どちらを好みますか？それはなぜですか？
[Do you prefer hot or cold weather? Why?]

APPENDIX I

EXTRAVERSION/INTROVERSION QUESTIONNAIRE ITEMS (ENGLISH VERSION)

1. Am the life of the party
2. Like to talk to a lot of people
3. Start conversations
4. Like to draw attention to myself
5. Like to be the center of attention
6. Talk a lot
7. Feel at ease in a large group of people
8. Am skilled at dealing with people

APPENDIX J
EXTRAVERSION/INTROVERSION QUESTIONNAIRE ITEMS (JAPANESE VERSION)

1. 盛り上げ役である。
2. いろいろな人と話すのが好きだ。
3. 人に自分から話し掛ける。
4. ほぼ誰とでも仲良くなれる。
5. 注目の的になるのが好きだ。
6. よくしゃべる。
7. たくさんの人の中にいると落ち着く。
8. 人と関するが上手だ。

APPENDIX K
WILLINGNESS TO COMMUNICATE QUESTIONNAIRE ITEMS (ENGLISH VERSION)

1. I’m willing to talk about things I like or I don’t like.
2. I’m willing to talk about my past experiences.
3. I’m willing to talk about my future plans.
4. I’m willing to give reasons to support my opinions.
5. I’m willing to disagree with other speakers.
6. I’m willing to ask questions about other people’s experiences.
7. I’m willing to be the first person to speak.
8. I’m willing to bring up a new topic.
9. I’m willing to suggest a change of topic.

APPENDIX L
WILLINGNESS TO COMMUNICATE QUESTIONNAIRE ITEMS
(JAPANESE VERSION)

1. 英語の授業の中に、英語でトピックの変更を提案する。
2. 英語の授業の中に、英語で他のスピーカーの経験について質問する。
3. 英語の授業の中に、英語で自分の将来の計画について説明する。
4. 英語の授業の中に、英語で新しいトピックを提案する。
5. 英語の授業の中に、英語でディスカッションで最初に発言する。
6. 英語の授業の中に、英語で他のスピーカーの意見に反対する。
7. 英語の授業の中に、英語で自分の過去の経験について話す。
8. 英語の授業の中に、英語で自分の意見をサポートする為に例を挙げる。
9. 英語の授業の中に、英語で自分が好きな事又は嫌いな事について話す。

APPENDIX M
SAMPLE TRANSCRIPTS OF MONOLOGUE SPEAKING TESTS

Note: For reading ease, new AS-units begin on a new line. Pauses are included and are measured in seconds. To protect student privacy, all names are pseudonyms.

1. Sho (high beginning class)

Monologue speaking test 1 (first 90 seconds):
hello (.79) I think (3.19) spending time with friends (.56) is (.83) more fun (.62) because (3.55) to (.34) to do (.98) other people (.62) is (.42) fun (.21) and (3.17) if I was alone (1.38) I don't (.45) I can't talk (.2) other people (7.3)
I (2.07) last year I (.78) I go to (1.18) I went to Disneyland with friends (4.22)
they are very fun (3.17)
so (5.69) I (.64) I often go to (1.77) game center (.34) but (8.97) go to alone is not (1.33) much fun (.3) but (5.53) go to with friends

Monologue speaking test 2 (first 90 seconds):
hello (2.42) I think (2.0) I prefer traveling (.5) in Japan (.28) because (1.75) to (.46) travel abroad is (3.83) use much (.72) many moneys (1.77)
but (.98) in Japan is (1.74) don't use money (1.88)
little (3.95) and (2.67) going abroad is (1.38) too (1.64) dangerous (2.89)
but (.5) many Japanese places (.83) are safety (1.42)
because I want to (3.1) go to (3.65) Japan (2.38)
I want to (.43) go to Hokkaido (.35) because I (.82) I have never been to (.37) Hokkaido (6.14)
Hokkaido have very (.76) have many danger dangerous delicious (.41) food

Monologue speaking test 3 (first 90 seconds):
hello (.87) I prefer (.41) watching (.35) tv (.33) more than (.92) playing computer games (.73)
because watching tv (.4) is (1.39) don't make (.63) much time (1.0)
because (.36) watching tv is (.32) only turning on (1.77) light light (2.16) turning on (.97) switch (1.03)
but playing computer game is (1.57) many have cord (1.97)
for example Playstation have (1.87) switch cord (.44) and (2.69) playing game disk (1.99)
playing (1.41) computer games has many time so I don't (1.55) do (.56) computer games much (1.75)
and watching (.65) tv has many (.37) kinds of tv programs (1.15)
for example variety (.68) quiz (.27) documentary (.26) news and so on (1.14)
I (1.22) watch (3.06) I often watch documentary (1.56)
I like Kanburia Kyuuden (.94) very much (1.94)

Monologue speaking test 4 (first 90 seconds):
hello (2.92) I (.47) I (1.29) I like (.82) watching sports more than playing sports (.76)
because (1.06) I don't (1.49) I am not good at (.67) playing sports (.89)
so (1.75) I watching (1.01) whatever so I don't like playing sports (1.72)
but (1.66) recently I want (.57) to (1.6) sometimes I see (1.58) watching baseball (3.4)
so I (3.04) I see (.66) Japan series in (1.43) last year (1.97)
Masahiro Tanaka is (.66) very strong (.27)
I (1.05) am so excited (2.53)
but (2.4) I (.67) have I don't have (1.02) a (.30) lot of information for baseball (1.06)
so (2.12) I studying (1.6) baseball's rule.

2. Saori (low intermediate class)

Monologue speaking test 1 (first 90 seconds):
I prefer winter than summer (.62)
it's mainly because (1.05) winter has many event (2.31)
for example (.51) Halloween and Christmas and (.44) new year's day and (.34)
birthday (.65) my birthday (1.66)
so very (.74) fun (2.02)
and (3.48) and the second reason (1.56) in (.36) in winter (1.14) if we feel (3.64) cool (.47) cool (1.05) we can (.3) we can wear many clothes (.27) so (2.93) so (1.52) we can (3.74) what do you say taeru (6.59) we can (5.23)
but (.68) summer (.66) summer (.83) in summer (.37) we don't (.61) we can't (1.5) take off (.53) clothes (1.28) so (1.59) winter (.35) is

Monologue speaking test 2 (first 90 seconds):
I prefer eating at home (1.13)
it's mainly because (.86) my mother is good at cooking (.42) so (.27) she (.3)
sometimes she (.95) she cook my favorite food (.95)
for example (.82) Okinomiyaki (1.1) or (1.07) Hayashi rice (1.41)
and (1.45) my mother comes from (.46) comes from is (.36) from (.8) Yamaguchi (.54)
so (2.56) she sometimes she (.5) cook (4.94) what do you say Kyoto (.44) Kyoto ryouri (3.95) Kyoto ryouri (2.58)
I like (1.62) it (2.66)
of course (.44) eating in a restaurant is very fun (1.14)
sometimes I (.76) I go to (.78) I go to restaurant (.4) for eating (.61) with friend (2.52)
but (.79) my mother's food is (1.67) is good (2.01) than

Monologue speaking test 3 (first 90 seconds):
I prefer traveling abroad than traveling in Japan (.79)
it's mainly because (1.71) in foreign country (.65) I can (.34) I can see many (.35) foreign (.54) foreign custom (.43) and (.97) history (1.37) in foreign country (1.2)
so I (.78) I want to go to (.3) foreign country (1.44)
and traveling abroad (1.65)

but (.82) I often go to travel in Japan (1.14)
it's mainly because (1.66) it (1.23) it (.25) took (1.31) little money (.27) so (1.16) I go to (.85) I often go to hot spring (.26)
and (.3) in Japan (.51) and (3.47) and Japanese (1.18) history place (.44)
for example Kyoto (.33) and hot spring place is Hakone (.59)
and Izu (.39)
we can go (2.73) to (.31) Hakone for one (1.6) one or two hours (.3) so (.36) to (.62) a near place (.98)
so (.35) I often go to

Monologue speaking test 4 (first 90 seconds):
I like (.53) I prefer (1.21) watching Tv (1.74) it's mainly because (.71) I don't use (1.16) computer games (4.18) but (3.82) I (2.15) after I entered the university (.47) it is no time to watch (.51) watch Tv (1.06) so (2.29) I (.85) I don't (.71) I (2.29) it is short time to watch Tv (.91) but I prefer (.72) watch tv (1.67) and (1.65) recently (.64) I (1.90) I often use (.57) Iphone (.55) so (.74) there are many (.32) application to (.70) play computer games (.32) so (1.98) sometimes I play (.47) computer games (1.29) but (2.86) prefer is watch Tv (1.9) and I like dramas (1.28) for example

3. Maya (intermediate class)

Monologue speaking test 1 (first 90 seconds):
I prefer to playing (.56) sports (.42) I it's mainly because (1.57) because I like (1.11) I like playing sports (.59) and (1.4) so I if (.8) I (.42) watching sports (.33) if I watch sport (.86) I can't play anything only watching (.76) but if I play some sports (3.2) I can enjoy the playing some sports (.62) so (.29) and (.29) playing when I play sports I can feel happy (.26) to (.48) happy (.69) to (1.19) when we when (.27) I won the game (.59) or success (.74) the (5.04) so (1.58) yeah I like playing (1.1) some sports and when I (.81) when I'm high school student (.9) I (1.24) I belong baton club baton twirling (.66) so now I when I watch the (.31) movie of baton twirling (1.03) I want to play baton twirling so much

Monologue speaking test 2 (first 90 seconds):
hi in my opinion I prefer birthdays (.54)
because I can eat cake (.94)
I like cake (1.07)
especially I like (.26) chocolate cake (.66)
so my mother (.92) sometimes my mother ordered birthday cake of chocolate (.9) for me so I'm happy to (1.45) I'm happy to eat chocolate cake with my family (.76)
and (.26) in supper supper (1.41) my mother made (.87) food which I (.5) I like so I'm happy to eat (.89)
then (1.15) I in (.5) birthdays (1.67) mother or father (.59) said happy birthday (.29) happy birthday Maya so (.58) I like (1.09) birthday's morning (1.24)
then but my birthday is April thirteenth *[date changed]* so (.48) in summer vacation (1.25) every year (.9)

so I can't (.43) I can't meet my friend (.75)
so I’m sad (.71)
but (.99) now I can use we can use (.67) email or phone (.27) or (.83) line (1.09) so (.38) every year.

Monologue speaking test 3 (first 90 seconds):
I prefer eating at home (.57)
because (1.24) we need to eat every day (.93)
so (1.2) if I eat at home (1.85) it is so healthy for me (.99) because (2.56) at (.92) eating at home we can (.97) cut (.85) the calorie or (.5) or (2.11) cut the (.41) oil (1.23) so (.61)
and we can choose (.99) we can choose (3.25) if (.6) if I want to (1.08) eat healthy food I can choose (.54) only vegetable (.7)
it is so easy at (.94) home (.85)
and (.98) eating at home is very (.27) cheap (1.76)
because if (.79) you (1.06) went to restaurant in Ikebukuro (2.04) maybe (4.16) at lunch (.34) I (.41) always (.59) paid one thousand yen or (.64) so (1.07) or more so (.78)
but (.43) at home (.25)

Monologue speaking test 4 (first 90 seconds):
hello my question is do you prefer traveling in Japan or abroad (.57)
and I prefer traveling in abroad (1.19)
it is because (2.36) I often saw Tv shows which introduce very beautiful (.53) place in abroad sightseeing place (.67)
so I want I have (.86) many (.29) place which I want to go (.64)
so (.89) I choose in abroad for example (.5) Tv shows introduce (1.12) white sand and (.35) blue seas (.28) place maybe in (.78) Australia (.28)
maybe (1.09) and (1.8) or (3.1) I want to (.55) I want to see the (1.34) very (5.64) no (.64) sorry (1.3) and (2.26) when I was a (.35) child we my family went to Hawaii (1.01)
and (.73) Hawaii's (.84) in winter (.84) so Hawaii is very (.86) hot (.37) in winter in

3. Sachi (high intermediate class)

Monologue speaking test 1 (first 90 seconds):
hi I prefer movies (.34) rather than Youtube (.47) because (1.42) movies we can enjoy (2.17) we can (.63) enjoy (.9) full (.53) big screen and (1.05) nice (.8) musics (1.05) like in (.2) Youtube (.46) it depends on (.37) the person's (1.32) computer spec like (.27) if the (.79) computer is (.28) old (.39) it (.56) works slowly so (1.37) the movies like slowly (.63)
and not good picture (.79)
but (.68) movies (.96) is (1.64) not depends on some (.42) those kinds of (.59) specs so (.33) we everyone enjoy (1.43)

and also (.84) like when we go in to (.68) theater (.37) theater and watch movie or like (1.58) event (2.1) this (.4)
we pay and (.57) the money will go to that (2.29) who create (.5) creators of the movie (1.22)
but Youtube is free so (.67) if you (.49) watch the (.93) movie or something like that we can't (.52) pay for it (.31)
and the creator is not (5.9) not

Monologue speaking test 2 (first 90 seconds):
I prefer watching sports rather than playing sports (1.68)
because (1.09) as watching sports (.4) when (.56) I (.57) sometimes there are a lot of (2.09) needs special abilities or (.51) maybe talent (.53)
but (.75) when (.42) I'm watching I can see (.26) a lot of those kind of talent (.61)
as sports so (.43) it's (.32) really fun to watch (.97)
and (.67) I (.67) I like watching (1.85) volleyball and figure skate (.49) figure skating and (1.04)
as figure skating there are a lot of (1.71) beautiful (2.71) techniques (1.14)
so (1.5) yeah it's really (1.25) fun to watch it and (1.09) really (.51) interesting (1.12)
in fact I (.35) can't play sports well so (.84) I don't like playing sports so
but (1.46) watching sports like soccer or (.31) baseball is (1.04) kind of (1.55) I don't like watching because like (3.76) like on (.4) like in the soccer there are a lot of (2.72) audience is

Monologue speaking test 3 (first 90 seconds):
I prefer (.34) winter (.33) better than (.44) better than summer (.3) because (.33) in winter we can (.71) because I love the (.34) relax time of (.6) end of the year to New Year (.97)
and (.73) because we can (.72) just (1.21) just enjoy and (2.1) New Year we can eat a lot of (.38) like (.36) delicious things (.29) like such as (.37) oseichi and (.51) ozoni or something like that (1.3)
then (1.08) in winter (1.8) if we can stay in the home we can use (.73) a heater (.58) heater (1.02)
so (.46) if we don't feel so too much cold (.27) or (2.74) too much cold so maybe it's (1.72) good to relax (.93) in winter (2.57)
in winter the (2.43) our (.36) bed (.8) is really warm and (1.8) warm and (.64) nice (.34) so (1) it's good to sleep (2.03)
because of the summer it's (.53) too hot and maybe (1.38) almost everywhere is (.26) not (.39) comfortable uncomfortable so maybe (.51) it's not good

Monologue speaking test 4 (first 90 seconds):
I (.47) prefer eating at home more than than in than in in a restaurant (.51)
because (.52) in the (.43) in the home we can (.35) eat (.69) we can eat more (2.63) like (.4) like vegetables and meats fish and (.4) some those kind of (.43) like more healthy (.58) foods (1.05)

and (1.07) one (1.08) one advantage could be we (.72) like (.33) children can help their mother (.64) maybe father (.63)
and (.57) they can communicate and also (.59) they children can learn how to cook from their (.27) their parents (1.14)
but (.49) another one (.31) disadvantage of (.48) of eating at home could be (.72) maybe it's (.59) it's like (.35) like every time their (.27) mother or father making their foods it’s kind of tired for them and (.47) like (.31) from (.36) making foods and also (.69) washing (.29) dish (.64)
so maybe just that one (.37) disadvantage of eating at home (.6)
but (.39) and eating at home we can (1.16) we can a lot of (.69) we can talk to (.27) talk with (.26) their (.68) talk with (.54) their (1.94) their (.4) with their family like (.89) through eating

APPENDIX N
SAMPLE TRANSCRIPTS OF 3/2/1 SPEAKING ACTIVITY

These transcripts are taken from lesson 8, the fifth data collection point of 3/2/1 speaking activities.

1. Sho (high beginning class)
3-minute speech
(.91) hello (.85) today (.43) we discuss (1.01) Japanese culture (1.98)
first (.69) Japanese pop culture is (1.64) popular in foreign (.42) people (2.54)
for example (1.46) manga anime karaoke (.27) and so on (2.49)
I (1.02) I like manga very much (.72)
so (1.05) I want to (.54) introduce (1.80) Japanese manga (1.24) for foreign people (4.22)
but (2.87) I interesting (2.51) Japanese (.58) traditional culture (.47) a little (6.50)
shogi (.72) shogi is (2.62) I (.77) want to learn shogi (6.24)
so (1.65) professional shogi player (1.27) character is very (2.02) unique (5.13)
for example Hashimoto (.73) Hashimoto (1.06) Nagata's character (.61) very (.89) interesting (.85)
so i want to learn (.70) traditional (.67) Japan traditional culture (4.34)
and (.46) but i want to (1.69) introduce to foreign (1.93) people (1.89)
i want to introduce Japanese pop culture because (3.63) Japanese pop (.42) traditional culture is (1.04) too difficult to learn (.88) for (.32) foreign (.61) people (4.57)
so i want to (.71) talk (1.38) each other (3.23)
my favorite (1.56) manga (4.05)
next (3.72) and (2.17) i want to learn (3.87) foreign culture (1.98)

2-minute speech
(.68) hello (3.15) today we discuss (1.12) about Japanese culture (1.81)
i think (.39) Japanese pop culture (.95) is very popular in foreign country (2.45)
manga anime (.29) karaoke (.80) and so on (.89)
i like manga very much (1.32) so (2.48) we (2.13) we (.75) we want to i want to discuss (.62) foreign country to (2.82) foreign (.46) people (1.01) to (1.82) my favorite comics (.70) manga (1.85)
but i want to (3.70) learn (.55) Japanese traditional culture (.88)
Shogi (.80) i want to learn Shogi (.26) because Shogi professional player's character (.35) is (.26) too unique (.66)
for example (2.80) Hashimoto Nagata's (1.23) joke is very (.90) interesting (2.39)
but (2.47) i (2.78) traditional Japanese (.34) culture (.92) is too (.34) difficult to introduce for foreign people (6.98)
but (1.33) but i (1.92) so i (.45) want to (3.39) do karaoke (.65) for foreign people (1.70)

1-minute speech
(.55) hello (.69) today we discuss (.59) Japanese culture (2.19)

I think Japanese pop culture (1.06) is (.32) very popular in foreign country (1.97)
so (.70) manga anime karaoke and so on (.79)
i want to (1.68) discuss (.51) foreign people to (1.45) my (.51) favorite manga (2.69)
but i (.42) i (.96) like Japanese traditional (.56) culture (3.10)
i (1.24) i like Shogi because (.49) Shogi's professional (1.24) player's character is (.25) very unique (.70)
for example Hashimoto Nagata's (.82) talk very jokes (1.08)

2. Saori (low intermediate class)
3-minute speech
(.32) today's class we discuss (.51) we talked about Japanese culture (.84) traditional culture or pop culture (1.93)
we discuss (4.32) we discuss (1.37) we talked about (.63) Japanese pop culture (.45) for example manga and anime (1.01)
manga and anime is very popular for foreign country foreign people (.96)
so (1.17) so foreign (1.85) American (1.94) anime and comic is very simple so (.87) battle (.25) so (.60) only battle (.91)
but Japanese (.30) anime and (1.09) manga (.76) is very high quality (.31) so (.69) it is very (.97) complicated (.43) so we (.49) interesting (.27) interested in manga and anime (1.30)
and we talked about traditional Japanese culture (.84)
for example (1.07) kimono (5.04)
we talked about (.63) kimono is (.47) very (.85) very stylish and (3.26) but (.74) it is (.62) not easy to wear (1.13) kimono (1.47)
and (.53) we don't (.49) we have not chance to wear kimono (1.66)
so (.90) it is (.40) unpopular (1.90) in young people (2.29)
and (1.11) tea ceremony (.57) show many Japanese culture (.84) for (.51) foreign (.45) people (2.71)
for example (.99) kimono and (1.68) sitting style and (1.45) tea (.73) and (1.77) Chagashi (2.22)
it is very (1.13) interesting Japanese style (.51)
so (1.24) we should introduce (.61) Japanese (.95) traditional culture for (1.62) foreign people (2.18)
and (.27) Nara and Kyoto have many traditional (.86) culture (.32) so (1.45) they (.60) they want to go to (1.71) there (1.49) and (2.06)

2-minute speech
(.64) today's class we talked about Japanese culture (.76)
we (.26) we discuss (.34) first (1.53) Japanese (.41) traditional (.33) culture (1.22)
we talked about (1.69) kimono and tea ceremony (2.46)
we (.75) we talked about that (.79) about (1.52) kimono is very (.58) stylish and (1.68) and (.26) but (1.12) but we don't we don't have chance to put on (.40) kimono (.94)
and (.60) it is not easy to (.33) wear (.95) kimono (.53)
so it is unpopular (2.14)
Yukata is unpopular in young people (1.72)

yes and (1.32) tea ceremony (1.51) tea ceremony show many Japanese culture (.80) for foreign (.48) people (1.43)
for example (.39) sitting style and (1.37) Yukata (.25) and (.59) Japanese (.65) rule manner (1.03)
so (2.97) so (3.44) many foreign people interested in (.34) tea ceremony (1.09)
and we discuss (.57) Japanese pop culture (4.37)
Japanese pop music is (1.0) is (.40) very popular (.62) in foreign country (2.08)
but we discuss (.90) we talked about (.37) Arashi

1-minute speech
(.82) for today's class we discuss (.78) Japanese culture (1.41)
first we talked about traditional Japanese culture (3.46)
for example tea ceremony and (.63) kimono (2.02)
we (2.23) kimono is very stylish and (.82) but (.38) we don't have a chance to put on (.41) and it is not easy to wear (.51)
and (.65) hair arrange (.94)
so (1.16) we are it is (.50) unpopular (2.53) kimono is unpopular in young people (.80) young people (1.75)
we (2.88) we have not (.26) chance to wear (1.06) and we discuss

3. Maya (intermediate class)
3-minute speech
(1.21) today's topic is (1.17) traditional culture Japanese culture (.89) like traditional culture or pop culture (.31)
in last discussion (.77) we talked about (1.96) pop which (1.51) which is important (.53) to (.59) important pop culture or traditional culture (.85)
and i think pop culture is (.80) more important (1.09) than traditional culture because (.45) pop culture (1.10)
i know traditional culture is dying so it (.38) it's need to protect (.46) but Japanese pop culture (.70) need to (.62) Japanese economy (1.21)
so Japanese pop culture is more important (1.26)
and (2.56) we talked about (1.15) which (1.56) which (.69) culture (.42) do you want to introduce to a foreigner (.44) friends (1.01)
then I think (1.32) pop culture or traditional culture then i think (1.0) pop culture is better because (.28) we can enjoy (.27) both (.87)
so (.93) pop culture i (.52) choose pop culture (.40)
then (.88) we talked about (2.51) J-pop J-movie karaoke print club (1.57) print club (.63) anime (1.44) manga (.95)
which is (1.41) interesting for foreigners (.35) or difficult for foreigners and i think (1.44) print club is interesting because when i (.65) when I (.93) go to America (.38) print club is (1.08) can't write message or (.64) can't use (1.18) stamps so i think (.29) print Japanese print club is so high quality (.76)
so it is interesting for foreigners (.60)
and difficult thing is J-movie because (.84)

for example (.60) Japanese love story is too difficult to (.61) Hollywood movies (.32) Hollywood Japan (.32) Hollywood (.52) love story (.54) like Japanese (.31) love story is more pure (1.02)
i think so (1.09)
maybe it is difficult to understand (.96)
then (1.47) today (5.49) we talked about

2-minute speech
(.81) hi today's topic is (.25) Japanese culture pop culture or traditional culture (.29) and in last discussion (.45) we talked about (.97) which (.76) which culture is important (.43) pop culture or traditional culture (.59)
then i think pop culture is (.80) important (.26) more important than traditional culture because (.78)
i know traditional culture is dying so they need to protect (.48) but (.43) I think Japanese pop culture (.36) need to (.27) Japanese economy (.88)
like (1.69) foreigners like manga Japanese manga or anime so they bought (.38) goods of (1.0) anime (.91) so that that is helpful (.52) to economy (.88)
then (2.35) we talked about (4.09) which culture do you want to introduce to a foreigner friends (.25) pop culture or traditional culture (.45)
then i (.87) in my opinion (1.63) pop (.50) i want to introduce pop culture to my (.25) to a foreigner friend because (.81) if (.49) i introduce traditional (.48) no pop culture (.83) we can enjoy both (1.08)
like take (.36) print club or (.36) listen (.74) Japan J-pop (1.0)
so I choose pop culture (.50)
and we talked about (2.30) which (.42) which (2.65) karaoke J-pop J-movie (.52) or other things which is (.64) are more interesting (.40) for foreigners (.40)
i think print club is most

1-minute speech
(.69) today's topic is culture Japanese culture
then in last discussion (.29) we talked about (.98) what is (.46) which is important (.42) pop culture or traditional culture
then i think (.30) pop culture is (.31) most (.82) more important (.43) because (.49) I know traditional is dying and so (.47) they need to protect but (.28) Japanese pop culture need to economics Japanese economics like (.46) foreigners bought anime to goods or comics goods or comic (.39) bought comic (.32) so that is helpful for economy (.79)
then (.29) we talked about (1.06) pop culture or traditional culture which is more (1.0) which culture do you want to introduce to foreigners (.49)
then i think pop culture i choose pop culture because (.75) we can enjoy (.28) with (1.88) enjoy with karaoke

4. Sachi (high intermediate class)
3-minute speech

(.35) hi (1.03) today i (.27) heard some (.61) ideas (.25) new ideas about (1.21) Japanese pop culture and (.90) traditional culture (1.91)
we discuss (.72) which is (.60) important (.47) to (.27) important (.38) Japanese (.26) pop culture or traditional culture (.73)
it was hard to (.42) define what is important means (.93)
so it takes time for us to choose topic (.81) but (.38) at first time (.46) we defined as (.30) economic (.45) economy (1.75) to (.27) but then (.95)
at first time we (.44) think pop culture is (1.22) important because (1.14) it's (.97) as in the textbook it says (.52) eighteen billion dollars (1.61) as (.67) like pop culture like manga and anime or something like that (1.24) so and also (2.18)
but as (.91) traditional culture (.37) we can (.27) really experience it like not (.27) through a internet (.85)
we can touch it and we can (.38) try it (.83)
we they they can try it (1.68) they can try it and (1.64) so they (.62) come to Japan and (.60) they (.27) pay (.43) pay money to (.46) experience this so it's (2.06) nice to (1.61) move the economy in Japan (1.11)
so (.46) then (.75) next time (.26) next (1.11) we define economy (.33) not economy important as (1.01) to (.65) which which is (.99) easy easy to introduce (.54) to (.67) young (.37) people (.43) in foreign country (1.21)
then (.53) we (.86) we think that (2.41) pop culture is (.46) good for (.41) first (.39) knowledge to (1.17) find people to (1.14) get understand (.33) get knowledge about (.31) Japan (.37) because (1.27) because (.44) it have (.46) some traditional (1.40) some (.32) traditional (.63) cultures like not only Japan but also (.90) other countries like (1.0) so (.88)
and also like (2.02) it's (.99) so they (.42) maybe they got easy understand (.68)
then (1.71) not (.44) only pop culture but (.41) traditional culture is (2.06) we want to (1.18) foreigners to know because

2-minute speech
(.29) hi (.68) we discussed what (.68) which is important to (.56) important pop (.29) Japanese pop culture or (.57) traditional culture (.72)
and (.60) so what at at first we (.36) tried to define what is important mean it it was kind of hard (.77)
so take it took time but (.31) at first we (.77) define it as (1.11) economic (1.33)
so (.58) then (.25) as a pop culture (.27) it's (.85) it's (.49) nice (.35) it's (1.33) not (.29) but it's (.328) it get profits like in the textbook it shows (.61)
it's (.44) cost not cost (.62) profit was (.26) eighteen billion dollars or something like that so (1.46) maybe pop culture is really (.72) moving economy (.34)
but (.71) other (1.07) traditional culture (.63) the (.38) foreigners come to Japan and (.34) pay money and (.70) try (.46) and experience that those kind of (.30) like (.48) tea ceremony and (.64) or (2.69) watching Kabuki or something like that so (.43) they can experience a lot of (1.20) traditional cultures and (1.31) through paying money so (2.88)
i forgot (1.55) as (2.12) not like pop culture they can (.89) they can (1.13) hear or watch it as through the internet (.78)

maybe (.61) low cost of maybe it's illegally but (.96) free thing and (.51)
but (.33) they (.67) ask (.50) traditional culture they need to pay so (.29) maybe it's (1.0) it's more nice to (1.16) nice for the economy

1-minute speech
hi (.77) we discussed what which is important to (.26) important pop culture or (.46) traditional culture and (.82) we it was (.31) hard to define what important means but (.52) at first we (.46) define it as (.56) economy (.36)
and (2.37) the pop culture (.48) is (1.47) good because like (.46) it's in textbook it says (.26) eighteen billion dollars is (.79) are (.26) earned from like profit (.80) as through the (.38) pop culture but (.33)
we discussed like (.81) traditional culture is (.50) really nice because we (.36) no not we (.27) the foreign (.45) people can experience it (1.58)
like like (1.32) i forgot (1.26) but (.31) they need to pay money to (2.20) try it

APPENDIX O
FLUENCY RATING SCALE

This scale was adapted from Nitta and Nakatsuhara (2014).

5	Speaks fairly fluently with only occasional hesitation, false starts and modification of attempted utterance. Speech is only slightly slower than that of a native speaker
4	Speaks more slowly than a native speaker due to hesitations and word-finding delays
3	A marked degree of hesitation due to word-finding delays or inability to phrase utterances easily
2	Speech is quite disfluent due to frequent and lengthy hesitations or false starts
1	Speech is so halting and fragmentary that conversation is impossible

APPENDIX P
ORAL INSTRUCTIONS PROVIDED TO FLUENCY RATERS

You will hear Japanese second language learners of English speaking on a variety of topics. All of the speech samples are two minutes long. I would like you to make a fluency judgment about each sample. Using the 5-point scale, give each speaker a rating based on their speaking fluency. This is the flow of the language – does the speaker have problems finding words, hesitating and pausing often, or do the words come easily? Don't worry about grammar mistakes – that doesn't matter. So a speaker who is very fluent – that is, the words just flow with no struggle, would be at the top of the scale, while someone who has a hard time expressing him or herself would be closer to the right end of the scale.

I would like you to try to use the whole scale over the course of the experiment. Please listen to the whole sample before making your decision.