A ONE-SEMESTER FORM-FOCUSED INTERVENTION ON THE DEVELOPMENT OF SPEAKING PROFICIENCY

A Dissertation
Submitted
to the Temple University Graduate Board

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

By
Chie Ogawa
August, 2019

Examining Committee Members

David Beglar, Advisory Chair, Teaching and Learning
Tomoko Nemoto, Teaching and Learning
Paul Leeming, External Member, Kindai University
Timothy Doe, External Member, Meiji University

ABSTRACT

This study was an exploration of the effects of a pedagogical intervention on the development of Japanese university students' oral performances. In task-based language teaching (TBLT), developing speaking proficiency is a major learning goal. However, research examining the effect of a focus on linguistic form in TBLT is limited. One way to balance communication and attention to linguistic form in TBLT is to add form-focused instruction to the communicative tasks. This study is an exploration of the longitudinal effects of form-focused instruction in a speaking task on the development of speaking proficiency. The current study was conducted for the following research purposes. The first purpose was to explore the longitudinal development of CALF (complexity, accuracy, lexis, and fluency) through form-focused intervention. A one-semester form-focused intervention was conducted to investigate how L2 learners develop or change their linguistic performance as measured by the CALF variables. The second purpose was to explore proceduralization through the 3/2/1 task. The third purpose was to investigate the relationship between communicative adequacy and CALF in the 3/2/1 task. This purpose was addressed by comparing human raters' perceptions of communicative adequacy with the CALF analyses. The final purpose was to qualitatively investigate what the participants prioritized during their task performances.

The participants were 48 first-year Japanese university students attending a private university in eastern Japan. A shortened version of the 4/3/2 task, the 3/2/1 task, was implemented 10 times for 13 weeks in one academic semester. In the 3/2/1 task, students talk about the same topic for 3 minutes, then 2 minutes, and finally 1 minute.

The participants were divided into three groups: the comparison group, the teacher-led group, and the teacher and peer group. Two types of form-focused instruction were implemented, teacher-led planning and a peer-check activity. The participants in the comparison group started the 3/2/1 speaking task immediately, those in the teacher-led group read a teacher-model passage with the target formulaic language underlined prior to beginning the 3/2/1 task, and those in the teacher and peer group received a peer-check treatment while doing the 3/2/1 task in addition to teacher-led planning. Listener partners checked to see if the speakers used the target formulaic language during the 3/2/1 task. The target forms were (a) stating opinions (e.g., *In my opinion*), (b) giving reasons (e.g., *It is mainly because…*), (c) giving examples (*For example…*), and (d) expressing possibilities (*If…*). Speaking data were collected at Time 1 (Week 2), Time 2 (Week 8), and Time 3 (Week 14), transcribed, and analyzed for syntactic complexity, morphosyntactic accuracy, lexical diversity, fluency and communicative adequacy.

This result showed that form-focused instruction with the target formulaic language improved the Japanese university students' speaking fluency such as mean length of run and phonation time ratio. The participants also improved human raters' perceptions of communicative adequacy over one academic semester. There was a significant and strong positive relationship between utterance fluency and human raters' evaluation of communicative adequacy. In addition, the peer-check enhanced the learners' usage of a wider variety of the target formulaic language.

The results indicated that including formulaic language instruction can enhance learners' mean length of run, which is a measure of speaking fluency, while teacher-led planning can help learners notice target forms. The peer-check can pressure learners to use the target forms during the 3/2/1 task and provide feedback so that speakers know what form should be used in the next 3/2/1 task performance. Suggestions for future studies regarding the use of formulaic language in TBLT tasks are proposed.

ACKNOWLEDGMENTS

First, I would like to thank my main dissertation advisor, Dr. David Beglar, for providing me support, insightful suggestions, and encouragement throughout my dissertation project. Since I entered the TUJ doctoral program, Dr. Beglar has taught me and guided me regarding study design, statistical analyses, and the writing process. I truly appreciate that he took his time to help me improve my dissertation study.

Second, I would like to express my gratitude to my dissertation defense committee: Dr. Timothy Doe, Dr. Paul Leeming and Dr. Tomoko Nemoto. Dr. Doe shared his ideas about his speaking research and helped me with the speaking analysis software PRAAT. Dr. Leeming provided me with helpful feedback on TBLT following my dissertation proposal defense, and Dr. Nemoto provided constructive and detailed feedback on my manuscript and statistical analyses.

I thank the human raters for taking their time to evaluate the participants' monologues during their busy schedule. I would also like to thank Takako Ogawa for checking my transcription of recorded speech data.

I would also like to thank the members of my doctoral cohort. I started the doctoral coursework at the Temple University Osaka campus and moved to the Tokyo campus. I would like to thank Nancy Lee and Yukie Saito-Oyama for sharing their ideas about speaking research by calling, skyping, and face-to-face meetings. I would also like to thank Michael Holsworth and Mayumi Asaba for their encouragement throughout the dissertation writing process. I truly believe that I was able to maintain my motivation with my peers' help and encouragement. I feel grateful that I could get to know the Osaka and Tokyo cohort members. I appreciate their support and help as well as their efforts to help each other improve academically. Last but not least, I would like to thank all of my students who participated in this study.

Chie Ogawa
Osaka, Japan
May, 2019

TABLE OF CONTENTS

APPENDICES

LIST OF TABLES

LIST OF FIGURES

Figure Page

CHAPTER 1

INTRODUCTION

The Background of the Issue

The ability to communicate is a crucial skill for humans to acquire. In addition to non-verbal communication such as facial expressions and eye contact, people use verbal communication to convey meanings. For example, verbal skills are used to express feelings, tell news, exchange information, ask questions, and discuss problems. With clear, coherent, convincing, well-organized speech, people can achieve many important goals, such as developing and maintaining human relationships, participating in discussions, getting jobs, or conducting business negotiations.

Communication skills in a foreign language have become increasingly important in today's era of globalization. In 2003, The Japanese Ministry of Education, Culture, Sports, Science, and Technology (MEXT) proposed an Action Plan to Cultivate *Japanese with English Abilities* to reform English education and improve Japanese students' communication abilities. The action plan is applicable to all levels of English education such as elementary schools, junior high schools, senior high schools, and universities. For example, from the 2011 academic year, MEXT started compulsory weekly classes called foreign language activities for Grade 5 and Grade 6 elementary school students (MEXT, 2009). Starting from 2020, MEXT will upgrade English to an official subject for fifth and sixth grade students, who will study three times per week. MEXT also began encouraging high school teachers to conduct English classes in English using more communicative activities in 2013, and they published the *Report on the future improvement and* *enhancement of English Education (Outline): Five recommendations on the English education reform plan responding to the rapid globalization* in 2014. In this proposal, due to ongoing globalization, the development of students' English proficiency is crucial for Japan's future. In addition, MEXT claimed that with an eye to the year 2020, in which the Tokyo Olympic and Paralympic Games will be held, the government will enact new English education reforms designed to enhance English communication skills. Verbal communication in English plays a crucial role in effective communication. With higher levels of English speaking proficiency, language learners can express thoughts, exchange ideas, and provide information more effectively when communicating with people from different language backgrounds.

Nowadays young people in Japan are expected to become global citizens, *kokusaijin* (MEXT, 2014). MEXT stated that Japanese people should expect to achieve top-level English proficiency in Asia. To function effectively as global citizens, people need to be able to use a wide variety of functional language to perform speech acts such as sharing ideas and stating opinions, yet, many Japanese university students have difficulty speaking English. One reason for the difficulty of speaking English is that they do not have sufficient opportunities to speak English in classroom. Most English classes in Japanese secondary schools are taught using the grammar-translation method (*yakudoku*), a method in which students rarely engage in speaking activities (Nishino & Watanabe, 2008). Approximately 55 % of Japanese high school students go on to university or junior college in 2018 (MEXT, 2018) and they spend a great deal of time studying English to pass university entrance examinations, most of which is likely to focus on reading comprehension (Nishino & Watanabe, 2008). Second, because most classes are taught using a teacher-centered style, many students have been accustomed to listening to their teacher in the classroom (Nishino & Watanabe, 2008); therefore, they have had few opportunities to express their opinions in the classroom. In sum, Japanese students need to be provided with opportunities to express their ideas and opinions in the classroom in order to acquire the communicative language skills that will allow them to work with people from different language backgrounds. Because time for foreign language instruction is limited, students in EFL contexts such as in Japan need to maximize their opportunities to practice speaking inside foreign language classrooms. This need is recognized by MEXT and many language-teaching professionals and as a result, MEXT has been encouraging teachers to use communicative tasks in foreign language classrooms (Nishino, 2011; Nishino & Watanabe, 2008; Tahira, 2012).

TBLT has been recognized as a teaching approach that emphasizes communication (e.g., Ellis, 2018; Long, & Crookes, 1992; Skehan, 1996, 2018). TBLT is goal-oriented, meaning-focused, student-centered, related to the real world, and needs-based (Ellis, 2003, 2009a; Long & Crookes, 1992; Samuda & Bygate, 2008; Van den Branden, Bygate, & Norris, 2009); thus, the primary focus of a task is to for learners to convey their meanings with a clear goal in mind and with a genuine need for communication. For example, in information gap tasks, two learners do not share the same information; therefore, they need to exchange the information to complete the tasks. Task-based language classrooms are student-centered, so learners must use their own linguistic and non-linguistic resources when performing the tasks (Ortega, 2012).

In conclusion, MEXT has taken a strong interest in the effects of globalization. The introduction of MEXT's Action Plan and the new course of study have influenced the teaching approach used in elementary schools, junior high schools, high schools, and universities. However, despite these positive changes, most Japanese students do not have enough opportunities to practice their speaking skills due to lack of opportunities to speak English outside of classroom and the lack of a student-centered curriculum. TBLT is one teaching approach that can increase students' opportunities to practice speaking.

Statement of the Problem

Researchers and teachers in communicative classrooms face several problems. The first problem is that it is not clear how L2 learners develop oral performances over an academic semester because few task-based researchers have explored the longitudinal development of learners' oral proficiency through task-based instruction and pre-task planning. In previous task-based studies, researchers have usually looked at participants' performances cross-sectionally to compare groups that did and did not engage in pre-task planning. Pre-task planning enhances learners' performances because it can decrease their cognitive load (e.g., Foster & Skehan, 1996; Ortega, 1999; Yuan & Ellis, 2003); however, language learning occurs gradually and it is therefore essential to understand how pre-task planning affects learners' oral performances over time, not merely in a one-time treatment.

The second problem is that it is not clear how L2 learners acquire target linguistic forms in a task-based classroom. It is sometimes said that the use of TBLT does not ensure that form-focused instruction occurs (Swan, 2005); therefore, the use of TBLT in Japanese educational settings has been questioned (Sato, 2010). Few researchers have examined empirically how L2 learners can acquire linguistic forms through tasks because

TBLT is usually meaning-focused (Bygate, Skehan, & Swain, 2001; Ellis, 2003; Prabhu, 1987; Van den Branden et al., 2009), and this tendency has led to misunderstandings among classroom teachers that TBLT involves limited instruction of linguistic form (Ellis, 2009a). According to Ellis (2016), there are several ways to incorporate form-focused instruction in TBLT: text enhancement, pre-task planning, corrective feedback, and repetition. Ortega (2012) also emphasized that pre-task planning plays an important role, as it provides teachers with opportunities to add form-focused instruction to a TBLT syllabus. Among the four types of form-focused instruction in TBLT, Ellis (2016) does not include explicit instruction in his list, which is one of the most straightforward ways to engage in form-focused instruction in a language classroom.

The third problem is that few researchers have examined how particular linguistic forms are proceduralized through the 3/2/1 task. In the 3/2/1 task, speakers talk about a topic for three minutes, then repeat the same content in two minutes and then in one minute. Although the 4/3/2 and 3/2/1 tasks have been investigated by several researchers (e.g., Boers, 2014; De Jong & Perfetti, 2011; Nation, 1989), it is uncertain what linguistic forms are proceduralized through these tasks because no intervention has been provided prior to or during the task. If form-focused instruction is incorporated into pre-task planning, learners are encouraged to use target forms communicatively (e.g., Ellis, 2016; Ortega, 2012). If they use the forms repeatedly in meaningful contexts, they might be able to acquire, proceduralize, and eventually automatize the target forms.

The fourth problem is that previous research on speaking tasks has primarily been focused on the quantification of complexity, accuracy, lexis, and fluency (CALF) in participants' performances; therefore, it is not clear if L2 learners can achieve a task goal communicatively. However, analyses of L2 learners' degree of communicative achievement is missing. Recorded speaking data are usually transcribed and analyzed manually by counting linguistic units such as clauses and error-free clauses. Recorded data can be analyzed using software to calculate the length of pauses and the numbers of syllables produced. While CALF measures provide a partial indication of the quality of learners' oral performances and help researchers track the development of oral proficiency, they do not indicate the degree to which students have achieved the communicative goals associated with the task. Few researchers have used human raters to assess students' communicative adequacy (e.g., Révész, Ekiert & Torgersen, 2016; Sato, 2011), even though the use of human raters can clarify the learners' degree of communicative adequacy as well as their command of the linguistic features measured by the CALF variables.

The fifth problem is that quantitative analyses do not always provide researchers with sufficient details concerning language acquisition. Previous task-based researchers have generally used quantitative analyses that do not show why learners' oral performances improved or what the learners focused on during task performances (e.g., Foster & Skehan, 1999). For example, even if learners improve their oral fluency or syntactic accuracy, researchers cannot know what strategies the learners used to achieve a good commend of the target linguistic features only by looking at the statistical results. Researchers must also gather qualitative data to better understand what the learners prioritize while speaking (e.g., meaning, linguistic form, or fluency). Gathering qualitative data through open-ended questionnaires and retrospective interviews can shed light on this issue.

Purposes and Significance of the Study

This study is motivated by four purposes. The first purpose is to explore the longitudinal development of Japanese ELF learners' speaking proficiency, and in particular, examine the effectiveness of a form-focused intervention on the learners' oral performance longitudinally. The intervention in this study is focused on teaching formulaic language. This purpose is significant because understanding how to incorporate form-focused instruction in communicative tasks is a necessary aspect of classroom instruction. In addition, including form-focused instruction in TBLT activities can lead to a better understanding of how form-focused instruction can be utilized as one part of communicative tasks (Ellis, 2005; 2009a). Some researchers have expressed concerns about TBLT because it does not provide learners many opportunities to acquire new linguistic forms (e.g., Sato, 2010; Swan, 2005). By identifying effective ways to include form-focused instruction, teachers can combine form-focused instruction with communicative tasks.

The second purpose is to explore the proceduralization of formulaic language through the 3/2/1 task. The 3/2/1 task was chosen because the task is considered effective for helping learners proceduralize their developing skills because of repetitive practice (De Jong & Perfetti, 2011). This issue is significant because the 3/2/1 task has been explored without a pedagogical intervention (e.g., Boers, 2014; De Jong & Perfetti, 2011; Thai & Boers, 2016). Previous researchers have suggested that learners should be provided with model input first and then encouraged to extract and use exemplars from the input in their own speech (Boers, 2014, p. 231). I decided to teach the formulaic language during the 3/2/1 task because the formulaic language can potentially help the learners speak more fluently because formulaic language allows learners to access prefabricated chunks that are stored in memory (Boers & Lindstromberg, 2012; Segalowitz, 2003; Wood, 2010, 2015). Therefore, including this pedagogical intervention can increase our understanding of the extent to which formulaic language can be proceduralized and automatized longitudinally through the 3/2/1 task.

The third purpose is to investigate the relationship between communicative adequacy and CALF in the 3/2/1 task. Tasks are usually goal-oriented, so the degree to which students achieve those goals should be assessed. In monologue tasks, such as the 3/2/1 speaking task, task goals can include stating opinions clearly, providing reasons for opinions, and elaborating on ideas coherently. However, previous researchers have only looked at learners' oral performances based on analytical CALF statistics and they have therefore not investigated the degree to which learners successfully achieved task goals. In order to help L2 learners develop the communication skills needed to succeed in performing an opinion-based monologue task, it is necessary to teach them how to achieve particular communicative goals. By incorporating human ratings in a task-based assessment, researchers can better understand the relationship between the extent to which students achieve a communicative goal and the objective CALF statistics.

The fourth purpose is to investigate what learners prioritize during task performance qualitatively using open-ended questionnaires and interviews. This issue is significant because few researchers have asked students how they planned, what they prioritized during their task performances, and how they perceived the task. One advantage of conducting mixed method studies is that they can provide stronger evidence for conclusions through the convergence and corroboration of quantitative and qualitative findings (Johnson & Onwuegbuzie, 2004). In order to better understand the quantitative results, open-ended retrospective questionnaires and follow-up interviews were used to gather data regarding the participants' perception of the 3/2/1 task, the pedagogical intervention, and what they prioritized before and during the 3/2/1 task. By analyzing the cognitive processes involved in task performances, researchers can understand how learners produced the utterances and why they chose them.

The Audience for the Study

Researchers, language test developers, and teachers are the primary audiences for this study. First, the findings of this study will be of interest to researchers involved in TBLT research. The majority of previous CALF studies of students' oral performances have been cross-sectional. In contrast, this study provides findings concerning the longitudinal development of students' oral proficiency; therefore, the findings of this study allow TBLT researchers to better understand the effects of focus on form interventions longitudinally.

Second, language test developers can benefit from this study in terms of the validation of rating scales used for assessing speaking proficiency. Major commercial standardized language proficiency tests such as the TOEFL iBT, IELTS, and TEAP use speaking tests that involve human raters, while SLA researchers have used objective CALF statistics to measure oral proficiency. Thus, there is still a gap between SLA research and language testing research in terms of measuring speaking proficiency. If language test developers better understand the relationship between CALF variables and human ratings, it would indicate which CALF constructs are more strongly correlated with human ratings. Therefore, the findings of this study are useful to those wishing to understand more about the relationship between CALF and human ratings and they can potentially be utilized for the future development of rubrics.

Third, the practical implications of this study can be of interest to second language teachers seeking to implement an effective communicative task into their classroom teaching. The main treatment in this study is form-focused instruction prior to the 3/2/1 task as input enhancement and during the 3/2/1 task as a peer-check activity. Empirical findings based on quantitative and qualitative analyses will help teachers determine effective ways to implement the 3/2/1 pedagogical intervention.

Delimitations

The first delimitation of this study is the sample group. The participants were first-year Japanese university students who generally had few opportunities to speak English inside or outside the classroom before entering the university. Therefore, the findings should be generalized with caution to L2 speakers in ESL contexts who have frequent opportunities to speak English outside of the classroom.

A second delimitation is the participants' oral proficiency. The participants' proficiency levels were from low intermediate to intermediate level (TOEIC 400-550). Although the TOEIC does not assess test-takers' speaking proficiency, it provides a general indication of the participants' receptive English proficiency. The participants generally were able to speak English to convey meaning but they were limited in terms of their productive vocabularies and syntactic constructions. The findings from the treatment might be generalizable to students at similar levels of oral proficiency.

The third delimitation is the genre and type of task, a monologic opinion-based speech, used in this study. This type of task requires that participants express personal opinions and tell a personal story. The topics used, such as fashion, studying English, media, and travelling, are familiar to the students. The tasks are monologic; therefore, the findings should not be transferred to other genres (e.g., picture retelling or decision making) and interactive tasks (e.g., dyadic tasks).

Organization of the Study

Chapter 2, Review of the Literature, consists of the following main sections: Levelt's model of language production, Complexity, Accuracy, Lexis, and Fluency (CALF) indices, assessing communicative adequacy, focus on form, explicit instruction in TBLT, text enhancement, peer feedback, pre-task planning, online planning, the Limited Attentional Capacity Model, proceduralization and automatization, Logan's Instance Theory, task repetition, Transfer Appropriate Processing (TAP), and formulaic language. At the end of the chapter, the gaps in the literature are described and the purposes of the study and research questions are presented. In Chapter 3, Methods, the participants, The English Language Curriculum and the Discussion Course, instrumentation, mixed method research design, procedures, data coding procedures, analysis, and Rasch analysis are described. In Chapter 4, Results, the results are presented, and in Chapter 5, Discussion, the findings are interpreted and the theoretical and pedagogical implications discussed. Finally, in Chapter 6, Conclusion, the findings are summarized, the limitations of the study are discussed, suggestions for future research are offered, and concluding comments are made.

CHAPTER 2

REVIEW OF THE LITERATURE

In this chapter, I review the literature on Levelt's model of language production, the CALF framework, assessing communicative adequacy, pre-task planning, focus on form, proceduralization and automaticity, and formulaic language. 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.

Levelt's Model of Language Production

Before explaining the CALF framework, it is essential to understand Levelt's (1989) model of language production because it is the most well accepted explanation of how information processing components work in speaking. Although Levelt's speech model was developed to describe L1 speaking processes, Skehan (2014) explained that "Levelt's Speech Model has to be the starting point for a credible analysis of the psycholinguistic processes involved in second language speaking" (p. 4). Levelt's speech model has been used by L2 researchers to understand speech production (see Izumi, 2003; Kormos, 2006; Skehan, 2009, 2018 for comprehensive explanations).

There are three main elements in Levelt's model of language production (Figure 1): the conceptualizer, the formulator, and the articulator. The conceptualizer is the first processing stage where speakers develop the ideas they wish to express. In the conceptualizer, speakers develop the propositional content of the message and decide what to say. For example, speakers select the relevant information, organize the information, and keep track of what was said previously (Muranoi, 2007). In order to do so, speakers access declarative knowledge of the content (i.e., encyclopaedic knowledge), the situation (i.e., situational knowledge), and how discourse is organized (i.e., discourse knowledge) (Towell, Hawkins, & Bazergui, 1996). The product of these mental activities is called a preverbal message.

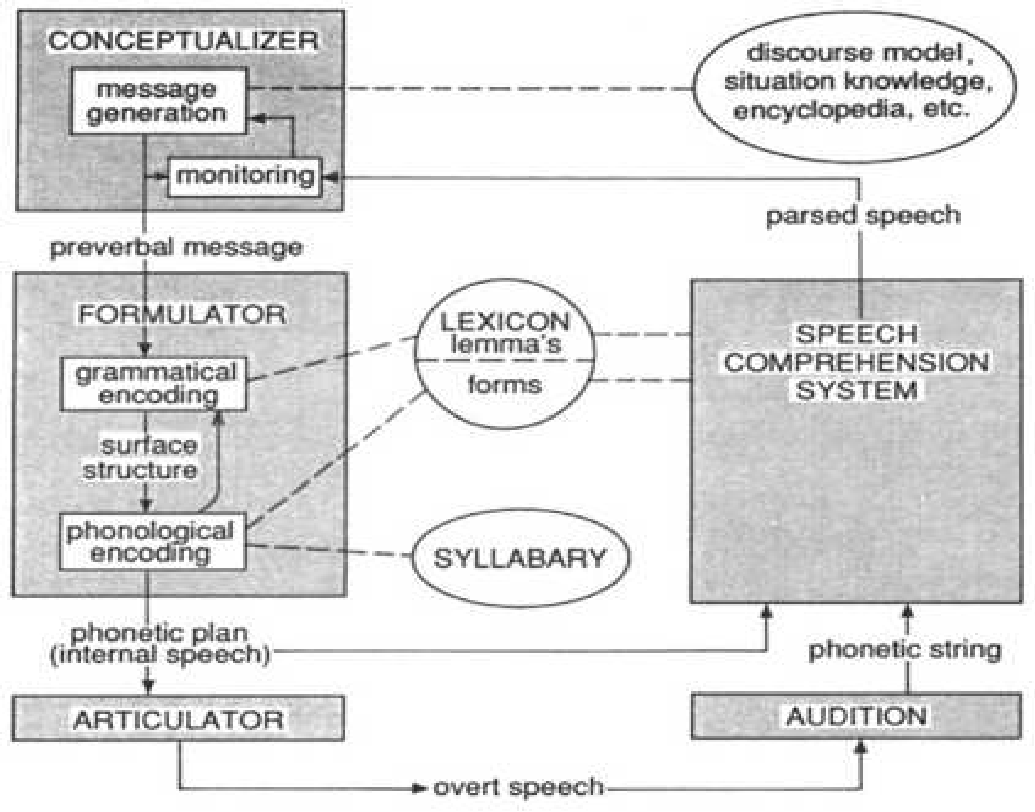


Figure 1. Levelt's blueprint for the speaker (Levelt, 1989).

The second stage is the formulator, where preverbal messages and ideas are transformed into language. In the formulator, speakers transform the preverbal message into a linguistic form, in which appropriate lemmas (i.e., knowledge of word meanings and the syntax associated with the lexis) are selected and grammatical and phonological rules are applied to create a speech plan. Lemmas are the word form that appears as a dictionary entry and that is used to represent all the other possible forms. For example, the lemma *give* includes *gives*, *giving*, *gave*, and *given*. The first step, grammatical encoding, includes procedures for accessing lemmas, which are form/meaning pairs contained in the lexicon. The lexical entry's meaning and syntax are represented in the lemma; lemma information is declarative knowledge that contains syntactic information about the lexical entry. In contrast, morphological and phonological properties are represented in the lexeme (De Bot, 1992, p. 2). The lexeme is the final form of a lemma. This process creates a set of surface syntactic forms and passes them to the phonological encoding part of the formulator. The second step, phonological encoding, includes retrieving or building detailed phonetic and articulatory plans for the lemmas. A phonetic plan is delivered from this formulator stage to the articulator.

The third step involves the articulator, which is where the speech plan is converted into spoken language. The phonetic plan created in the formulator stage is temporarily stored and fed back to the speech-comprehension plan and sent to the articulator. The phonetic plan consists of syllable programs so that the speaker does not need to generate or invent syllables from scratch. According to Levelt's model, monitoring is involved throughout the speech processing model. For example, speakers can monitor internally what they would like to say and then they link that information to the preverbal message by accessing the mental lexicon. Speakers have an internal model of their own speech system to produce particular sounds. The quality of monitoring directly concerns the quality of speaking performance especially in terms of morphosyntactic accuracy (Skehan, 2014).

Levelt's model of language production represents the language production of adult monolingual native speakers, yet many TBLT researchers have drawn on Levelt's model to explain L2 speakers' speech processing (e.g., Ahmadian & Tavakoli, 2011; Kormos, 2006; Kormos & Trebits, 2012; Maad, 2010; Skehan, 2018; Towell et al., 1996). De Bot (1992, 1996) adapted Levelt's model for bilingual processing. However, De Bot (1992) proposed that Levelt's idea of the conceptualizer had to be modified for bilingual speakers. While Levelt explained that the conceptualizer is language-specific, De Bot argued that the knowledge component is not language specific and that the conceptualizer is partly language-specific and partly language-independent.

The conceptualizer has two parts: macroplanning and microplanning (De Bot, 1992). In macroplanning, the language to be used is selected on the basis of information from the discourse model (p. 8). The discourse can differ depending on the speaker's L1. For example, De Bot (1992) gave examples of special references for Dutch, which only makes one conceptual distinction (e.g., proximal/distal: here/there), while Spanish makes three distinctions (e.g., proximal/medial/distal). Microplanning is necessarily language-specific because this is where speakers encode the preverbal messages.

De Bot (1992) also explained that in the formulator, the preverbal message is converted into a speech plan by bilingual speakers; however, De Bot argued that cross-linguistic influences have to be accounted for in the formulator stage (p. 6). The link between meaning and syntactic information contained in the lemma is a key aspect of the formulator. De Bot suggested that bilingual speakers have one lexicon in which different languages are stored together. This hypothesis provides a good explanation of code-switching and the storage and retrieval of lexical items. For example, one common problem that bilingual speakers face is that they cannot find relevant words to express a concept in a particular language. When L2 knowledge is insufficient, speakers might borrow from the L1, a strategy that allows them to produce utterances.

De Bot (1992) explained that bilingual speakers must have one articulator that has an extensive set of sounds and pitch patterns from both languages to work with (p. 17). There is evidence of cross-linguistic influences on pronunciation and prosodic patterns (De Bot, 1992); thus, L1-models of pronunciation and phonological encoding continue to play a role in L2 oral production. For example, successive bilinguals have many L1 intonational or prosodic characteristics in their L2 (p. 17). For fluent L2 speakers, all three stages are highly automatized; they involve a self-monitoring process and they operate in parallel (Skehan, 2014). On the other hand, non-native speakers' processing in the formulator is more effortful, and it includes repairs and repetition.

Ellis (2009b) stated that pre-task planning provides a theoretical account of learners' L2 performances, as it allows learners to attend to all three components in Levelt's model. In other words, pre-task planning assists conceptualization, which contributes to greater fluency and complexity (Ellis, 2009b). The effects of pre-task planning can also impact the formulator and the articulator because planning allows learners to access linguistic sources, provided that their attention is directed toward them. Unlike native speakers, whose formulator and articulator stages are highly automatized, L2 speakers are likely to be influenced by external factors such as repetition or pre-task planning because these interventions can allow them to pay more attention the linguistic form in the formulator (Muranoi, 2007; Towell et al., 1996). Therefore, Muranoi argued that pedagogic interventions can promote proceduralization in the formulator.

Skehan (2009, 2014, 2018) made a strong connection between Levelt's speech model and speech processing in task-based speaking performances. Skehan (2009, 2018) explained that native speakers can engage in parallel processing (e.g., the formulator deals with the previous conceptualizer cycles while the conceptualizer simultaneously attends to the next cycle) because their mental lexicons are extensive and well-organized. On the other hand, non-native English speakers take time to formulate how to produce utterances in English because it takes them more time to retrieve accurate linguistic forms in the formulator stage. At the same time, the basis for more accurate speech in second language speakers arises in the formulation stage, which makes them use more effort in the formulation stage to pay attention to lemma retrieval and syntax building. This idea suggests that second language learners are limited in terms of what they can focus on during meaning-oriented communication.

In sum, Levelt's speech model is a widely accepted L1 speech production model. Skehan's Limited Attentional Capacity Theory hypothesizes that L2 speakers usually cannot process linguistic information automatically in the formulator stage and they therefore take time to formulate what they have created in the conceptualizer stage. It is important to research on how pedagogical intervention helps L2 learners become more automatized speakers. Pedagogical intervention such as pre-task planning can help L2 learners to speak more fluently because interventions can potentially allow them to access their linguistic resources before engaging in the task.

Complexity, Accuracy, Lexis, and Fluency (CALF) Indices

CAF research started in the 1970s (Hosen, Kuiken, & Vedder, 2012). Researchers initially investigated L2 pedagogy by analyzing fluent L2 speech and accurate L2 usage in order to describe communicative L2 proficiency in classroom contexts. Skehan (1996, 1998) introduced a speaking proficiency model for the first time using the terms complexity, accuracy, and fluency. These concepts appeared often together as indicators of L2 learners' performances in investigations of the effects of other factors such as L2 attainment, the effects of instruction, and individual differences. Recently, the CAF components have played a central role in their own right due to the cognitive turn in L2 research (e.g., DeKeyser, 1998; Larsen-Freeman, 2006). These three CAF dimensions have been identified as distinct areas of L2 performances based on factor analyses (Housen, Kuiken, & Vedder, 2012; Norris & Ortega, 2009). One advantage of using the CAF indices is that researchers can capture L2 learners' performance and proficiency comprehensively because L2 performances are multi-componential in nature as shown by the notions of complexity, accuracy and fluency.

The CALF indices have been used as an indicator of learners' oral proficiency and language acquisition (Housen et al., 2012). Moreover, some researchers have used human raters to measure each CALF component (e.g., Iwashita, McNamara, & Elder 2001; Nitta & Nakatsuhara, 2014); however, the majority of task-based researchers have employed analytical measures, in which they analyzed transcribed speech data objectively using the CALF indices.

According to Housen et al. (2012), complexity refers to "the degree of elaboration, the size, breadth, width or richness of the learner's L2 system or repertoire"

(p. 25). Ellis (2003) similarly defined complexity as “the extent to which the language produced in performing a task is elaborated and varied” (p. 340). Complexity can be defined as the *breadth* and *depth* of L2 structures (Housen et al., 2012, p. 27). *Breadth* concerns grammatical and lexical diversity, while *depth* of L2 structure has to do with grammatical/lexical sophistication, that is, the embeddedness or compositionality of L2 structure. Although Housen et al. proposed different categories of complexity, they acknowledged that these categories are closely intertwined.

Syntactic complexity in speaking research has generally been measured using speech units such as the Analysis-of-speech Unit (AS-unit) (Foster, Tonkyn, & Wigglesworth, 2000). Syntactic complexity is commonly measured by calculating sentence length (e.g., mean length of AS-units) (Norris & Ortega, 2009). A longer AS-unit indicates a speaker’s ability to produce more complex utterances. Researchers also usually calculate the number of words produced in an AS-unit. For example, *I usually enjoy watching baseball at home rather than playing* (10 words) can be considered more complex than *I enjoy baseball* (3 words).

Another way to measure syntactic complexity is based on subordination (i.e., how many clauses are in an AS-unit) (Norris & Ortega, 2009). More subordinate clauses in an AS-unit means that it is a more complex utterance. For example, producing *I enjoyed the baseball game* <u>because</u> *my brother was in the team* can be considered more complex than *I enjoyed baseball in high school* because the first sentence has two clauses while the second has one clause.

The final way to measure syntactic complexity is based on the frequency of use of certain grammatical forms, which is a measure of sophistication. Some researchers (e.g., Ellis & Yuan, 2005; Robinson, 2007) have counted the raw frequency of certain syntactic forms such as passives, tensed forms, or auxiliaries. Norris and Ortega (2009) stated that this measure is used only in the field of SLA. One disadvantage of this measure is that different researchers analyze the same linguistic features using different criteria. For example, the 3rd person singular present *-s* form was characterized as a simple feature by Krashen (1994), and a complex feature by DeKeyser (1998).

In addition to syntactic complexity, more researchers (e.g., Skehan, 2009; Thai & Boers, 2016) are including analyses of lexical complexity in assessments of oral proficiency. Recently, the importance of lexis has been emphasized in an effort to better understand the psycholinguistics of second language speech production and the inter-relationships among the CALF components (Skehan, 2009, p. 512). Lexical complexity is important because the lexis-syntax connection is vital in performance areas and vocabulary is strongly associated with fluency (Skehan, 2009, p. 514).

Lexical measures are generally categorized as lexical sophistication and lexical diversity (Skehan, 2009). Lexical sophistication is defined as the percentage of sophisticated or advanced words in a text. Low-frequency words are generally considered sophisticated (Laufer & Nation, 1999). For the analysis of spoken language, Lambda is often employed. Lambda which is calculated by dividing a text into 10-word chunks and counting the number of difficult words in each chunk, represents the best fit of the distribution of the number of difficult words (Skehan, 2009, p. 515). Other measures of lexical sophistication are Laufer and Nation's Lexical Frequency Profile (1999) and measures calculated from Tom Cobb's Lexical Tutor web site (www.lextutor.ca).

Lexical diversity refers to the range and variety of vocabulary deployed in a text (McCarthy & Jarvis, 2010). Lexical diversity can be determined by calculating the Type-Token Ratio (TTR), in which the number of different words a L2 learner produces is divided by the total number of words. However, one disadvantage of the Type-Token Ratio is that it is strongly influenced by text length. Recently, other alternative measures such as D (Malvern & Richards, 2002) or the measure of textual lexical diversity (MTLD; McCarthy & Jarvis, 2010) are preferred. Both D and MTLD can assess lexical diversity without the influence of text length (McCarthy & Jarvis, 2010). According to Koizumi (2012), compared to TTR or D, MTLD is the least sensitive to text length when the text consists of at least 100 tokens.

Accuracy is related to target language norms. As such, accuracy refers to “the extent to which an L2 learner’s performance deviates from a norm” (Housen et al., 2012, p. 4). Housen et al. presented a wider definition of accuracy as appropriateness and acceptability because the norm differs depending on the social context. Having said that, the determination of appropriateness or pragmatically correct language might be difficult for low-intermediate-proficiency learners in a non-English speaking country; thus, I define accuracy as morphosyntactic accuracy in this study.

Morphosyntactic accuracy can be categorized as either global accuracy or specific accuracy (Iwashita, Brown, McNamara, & O'Hagan, 2008). Researchers identify all types of morphosyntactic errors when assessing global accuracy (e.g., Foster & Skehan, 1999). One advantage of measuring global accuracy is that it is comprehensive; thus, it provides a general understanding of the accuracy of learners’ utterances. However, one disadvantage is that it might be difficult for low-intermediate learners to produce error-free clauses or error-free AS units in their oral performances (Gunnarsson, 2012; Norris & Ortega, 2009; Pallotti, 2009). Another disadvantage is that global accuracy is too vague to capture what types of linguistic errors are made. For example, global accuracy measures do not indicate whether the error exerts a minor or major influence on the comprehensibility of the utterance.

Specific accuracy concerns target morphosyntactic features, such as present and past tense verb morphology (Gunnarsson, 2012). Mochizuki and Ortega (2008) explored learners' use of relative clauses because they were the target linguistic features in their treatment. One advantage of specific accuracy measures is that they are sensitive to treatment effects. On the other hand, one disadvantage is that they do not indicate the learners' overall level of accuracy.

Fluency historically refers to the smoothness of speech or native-likeness of speech (Housen et al., 2012, p. 4); however, fluency is considered in a narrower sense in Applied Linguistics (Lennon, 1990; Tavakoli & Skehan, 2005; Segalowitz, 2010). In this narrower sense, fluency refers to the speedy and smooth delivery of speech without pauses, repetitions, or repairs (De Jong, Hulstijn, Schoonen, & Groenhout, 2015). Based on Levelt's model (1989) and De Bot's (1992) interpretation of that model, L2 speakers are disfluent because they are in the process of developing lexical and grammatical knowledge and skills. Due to L2 speakers' slower processing during the formulation and the articulation stage, they are more likely to be disfluent. According to Segalowitz (2010), this particular feature of L2 learners' processing is called *cognitive fluency*, which is the speaker's ability to translate their thoughts into speech smoothly. However, cognitive fluency is difficult to measure because researchers cannot see inside the learners' brain, therefore, *utterance fluency* is used to understand processing difficulties by measuring particular aspects of L2 speech such as pauses or the syllables produced in a given amount of time.

Three subdimensions are recognized in utterance fluency: speed fluency, breakdown fluency, and repair fluency (Tavakoli & Skehan, 2005). Speed fluency, which refers to the speed or density of linguistic units, is usually measured by speech rate (e.g., number of syllables per minute). Breakdown fluency is the number, length, and location of pauses (Housen et al., 2012, p. 5). It is usually identified by measuring pauses or identifying where pauses occur in an utterance (e.g., pauses at the end of a clause or in the middle of a clause). Repair fluency refers to how often speakers use false starts, self-correct, or make repetitions.

De Jong and her colleagues (e.g., De Jong, Steinel, Florijn, Schoonen, & Hulstijn, 2013) have raised researchers' awareness of the importance of measuring utterance fluency more precisely. For example, speed fluency should be measured by calculating a measure of speed such as the number of syllables per second. Some researchers have used global measures of fluency such as syllables per total time including pauses. Including pausing time (e.g., syllable per minutes), however, does not accurately show how fast a speaker speaks because combining the speed of speech and pauses leads to confounded measures. For example, the measures of speech rate (number of syllables divided by total time including silences) and mean duration of a silent pauses both depend on the duration of silent pauses in the speech; therefore, these two measures are interrelated. If the two measures are strongly related to the fluency gain, the relative contribution of each measure is unclear due to multicollinearity of the measures (Bosker, Pinget, Quené, Sanders, & De Jong, 2013). De Jong et al. reported the correlation between these fluency measures and argued that using measures with low correlations with one another aids the interpretability of results.

The last type of fluency, perceived fluency (Segalowitz, 2010), refers to the listener's impression of the speaker's fluency. In this regard, perceived fluency is the perception of how easily and efficiently the listener was able to listen to the speech.

Many researchers have pointed out problems regarding CALF. The first issue is that CALF is a complex system and the optimal way to analyze oral data is unclear. Norris and Ortega (2009) stated that complexity, accuracy, and fluency are each complex subsystems with multiple parts. For instance, even though one component such as syntactic complexity can be measured in different ways, such as sentence or clause length, amount of subordination, or the variety of lexical or syntactic forms produced, few previous researchers have measured complexity in multiple ways. Moreover, some complexity measurements are redundant, as they measure the same construct. For instance, mean length of utterance and mean length of clause are both measures of length, while the mean number of clauses per AS-unit and the mean number of subordinate clauses per total clauses are measures of subordination. Norris and Ortega (2009) suggested that researchers think of why selected measurements are used and adopt more sustainable practices to contribute maximally to a replicable and cumulative understanding of the CALF indices across study contexts. It is essential for researchers to include reporting practices such as the accurate and complete description of measurement tools, data, and analysis.

The second issue concerns the adequacy of CALF measurement. Pallotti (2009) pointed out that few researchers have discussed the successful achievement of communicative goals even though most researchers have used communicative tasks. Pallotti defined adequacy as “the degree to which a learners’ performance is more or less successful in achievement the task’s goals efficiently” (p. 596). She presented the sentence *Colorless green ideas sleep furiously on the justification where phonemes like to plead vessels for diminishing our temperature* (p. 596) and stated that this sentence might produce high CALF statistics while being irrelevant to task success. In contrast, an utterance with low CALF statistics might successfully achieve the communicative task goal. This problem occurs due to the assumption that “more is better” (p. 597) where complexity, accuracy, and fluency are concerned; however, higher levels of complexity, for example, do not necessarily result in better performances. By analyzing to what extent learners achieve successful oral performances, researchers can better understand the quality of the learners’ production in terms of communicative goals both as an independent construct based on task success and as a way of interpreting CALF measurements (Pallotti, 2009, p. 599). Thus, Pallotti suggested that functional adequacy is a separate dimension that exists alongside CALF measurements.

To summarize, CALF indices have been widely used to analyze L2 speakers’ performances and language development in the field of task-based research. Complexity and accuracy are related to learners’ explicit and implicit procedural knowledge. In Levelt’s (1989) speech production model, complexity and accuracy are related to the conceptualizer and the formulator, while fluency is related to the speed and efficiency of the learners’ linguistic processing system and the degree to which learners have proceduralized their declarative knowledge and automatized the processes that make up Levelt's formulator and articulator (Housen et al., 2012, p. 6). L2 learners' performances are multi-componential, as they include constructs such as complexity, accuracy, and fluency. Thus, CAF indices allow researchers to capture important aspects of L2 learners' performance and proficiency; however, there are problems with using only CALF indices given that they do not always accurately indicate the degree to which L2 learners' demonstrate communicative adequacy during task completion. Therefore, task-based research is now shifting toward assessing communicative adequacy in addition to the analytical CALF measures.

Assessing Communicative Adequacy

As mentioned in the previous section, quantifiable CALF measures do not reliably indicate to what extent speakers achieve a communicative goal. Pallotti defined communicative adequacy as "the degree to which a learners' performance is more or less successful in achieving the task's goals efficiently" (p. 596). I define it more narrowly in this study as the degree to which a learners' performance is successful in achieving the task's goals in terms of monologue organization and linguistic competence.

Many TBLT researchers have assessed learners' performances based on CALF measures, but they have disregard their overall communicative effectiveness. This approach is note ideal because speakers who use simple and inaccurate linguistic forms can sometimes successfully achieve a task goal, while speakers with high levels of complexity, accuracy, or fluency might not be successful in functional terms (Révész et al., 2016). For this reason, only using CALF measures is insufficient when assessing task-based performances (De Jong et al., 2012; Ortega, 2003). Pallotti (2009) proposed that adequacy should be a separate measure of oral proficiency, independent from the CALF measures. In real-life settings such as classroom or test settings, the degree to which classroom learners or test-takers can function successfully needs to be given considerable weight (Révész et al., 2016). In order to better assess learners' communicative achievement, a subjective rating scale should be used in addition to objective CALF measurements (e.g., de Jong et al., 2012; Iwashita et al., 2008).

Some testing researchers (e.g., Iwashita et al., 2008; McNamara, 1990) have used human raters to assess learners' oral performances; however, in many cases, these rating scales are based on linguistic features such as grammar (e.g., grammatical accuracy and complexity), phonology (e.g., pronunciation, intonation, rhythm), and fluency (filled and unfilled pauses, repair, total pausing time, speech rate) (e.g., Iwashita et al., 2008; Nakatsuhara, 2012). Human ratings based on these criteria are likely to assess learners' command of linguistic features rather than the degree of communicative adequacy.

Some proficiency tests include rating scales that are used to assess both linguistic and non-linguistic features. For example, the TOEFL iBT evaluates the test-takers' ability to elaborate monologue speech through three categories: delivery, language use, and topic development (Educational Testing Service, 2014). Delivery and language use are linguistic features and topic development is a non-linguistic feature. For instance, delivery concerns fluency, fluidness, intonation, and pronunciation, while language use is related to the degree to which test-takers use accurate lexis and morphosyntax and a wide range of expressions. Topic development concerns coherence and the elaboration of speech (Educational Testing Service, 2014).

The TOEFL iBT speaking section includes six speaking tasks: Two independent tasks and four integrated tasks. In the independent tasks, test-takers talk about given questions within the allotted time, while, in the integrated tasks, test-takers receive input first by reading a passage or listening to a conversation. They then discuss the given topic. The independent tasks are monologue tasks. Both linguistic features such as delivery and language use, and non-linguistic features such as topic development are potentially important because the maturity of ideas and content development are aspects of successful communication (Sato, 2011).

An important question concerns what criteria to add to linguistic features in order to assess overall communicative adequacy. Révész et al. (2016) examined speech samples based on adequacy and a range of CAF indices. The researchers found that a set of linguistic factors significantly impacted communicative adequacy as perceived by trained raters. Specifically, the frequency of filled pauses and breakdown fluency were the strongest predictors, with fluency emerging as a critical determinant of communicative adequacy. In addition, other measures such as complexity (syntactic complexity, subordination complexity), accuracy (general accuracy, connector accuracy), and other fluency measures (silent pause frequency and speed fluency) had significant but weaker relationships with communicative adequacy.

Further research in this area was conducted by Sato (2012), who examined the relative contributions of linguistic criteria and speech content to scores on a speaking proficiency test. He was concerned that learners' ideas are emphasized in English-for-academic-purposes (EAP) courses but not in general English oral proficiency assessment. Sato defined speech content as topic development in the TOEFL iBT speaking rubric because test-takers try to convey relevant and well-elaborated ideas on the given topics. Nine raters assessed 30 students' monologues on three prompts. The test-takers had 30 seconds to prepare and one minute to deliver their speech. First, raters listened to each monologue to assess overall communicative effectiveness on a five-point scale ranging from 1 (*Unsatisfactory*) to 5 (*Completely satisfactory*) based on their intuitive judgment. The raters assessed overall communicative effectiveness but did so without a detailed definition. Sato explained that no detailed rating rubrics were used because the purpose of the study was to gather the raters' intuitive judgments of the test-takers' performances. After the raters assigned overall communicative effectiveness scores to the monologues, they listened to the monologues again and scored them using the following criteria: Grammatical accuracy, fluency, vocabulary range, pronunciation, and content elaboration/development. Unlike overall communicative effectiveness, the definition of each component was given to the raters. The ratings were analyzed using Rasch measurement, multiple regression, and multivariate G-theory.

Sato's (2011) findings are noteworthy because this is one of the few examinations of the extent to which speech content affects perceived proficiency. The results showed that pronunciation had the weakest relationship with overall communicative effectiveness, while content elaboration/development and fluency had the highest correlation. Standardized regression coefficients showed that content elaboration/development was the strongest predictor (β = .42), followed by fluency (β = .25). This finding is significant because speech content (e.g., topic development) is a crucial component of oral performances in academic settings (p. 235). These findings as well as those from previous empirical studies (e.g., Révész et al., 2016; Sato, 2011) indicate that CALF measures are somewhat associated with communicative adequacy.

To summarize, CALF indices do not adequately measure L2 learners' communicative achievement. In addition to the CALF indices, human raters' perceptions of the performances are necessary. Researchers have measured communicative adequacy by asking human raters to rate L2 speakers' performances using overall communicative adequacy holistic scores. These communicative adequacy scores have been found to be related to fluency (e.g., Révész et al., 2016; Sato, 2011) and content elaboration/development (Sato, 2011). However, because only a handful of researchers have examined communicative adequacy with respect to CALF measures, more studies should be conducted with different types of tasks and in different educational contexts.

Focus on Form

In TBLT, in which meaning is a primary focus, learners engage in various types of oral communication tasks, such as retelling tasks, information gap tasks, and decision-making tasks. In spite of the movement toward communicative approaches in foreign language instruction, there is still a misunderstanding among some language teachers regarding form-focused instruction (Ellis, 2009a). For example, some teachers have expressed concerns that TBLT cannot ensure adequate coverage of grammar and that attention to form in TBLT is limited (Swan, 2005). Indeed, previous studies in French immersion programs have shown that a teaching approach with a focus on meaning with no focus on grammatical form did not guarantee learners' development of grammatical accuracy or morphological features (e.g., Lapkin, Hart, & Swain, 1991).

One reason for this lack of acquisition is provided by Schmidt's Noticing Hypothesis (1990), which states that noticing or drawing learners' attention to target linguistic forms is an essential condition for L2 learning. This idea became a challenge to proponents of non-interventionist approaches because they believe that the only way to acquire a language is through natural exposure to the target language. Since then, many researchers have begun to investigate how form-focused instruction contributes to language development. Nassaji (2016) summarized this historical change as a shift in "focus from whether FFI (Form Focus Instruction) has an effect to what type of FFI is most beneficial" (p. 36).

Long (1991) introduced the distinction between focus on forms (FonFs) and focus on form (FonF). FonFs is defined as a teaching approach in which language is presented isolated from a meaningful context. In contrast, FonF is defined as the a approach that draws learners' attention to certain linguistic forms in the context of meaning-focused communication. Long's introduction of FonF has become the impetus for many recent studies that have attempted to explore the best way of drawing learners' attention to form and its effects on language learning. Given that FonF needs to occur in communicative contexts, it requires the use of tasks that focus learners' primary attention on meaning but also provide periodic attention to form by the teacher and/or students when this is triggered by communicative need. Therefore, Ellis (2016) stated, "Focus on form is a central construct in TBLT" (p. 405).

There are various ways to add form-focused instruction to communicative tasks. Willis (1996) argued that attention to form should be restricted to post-tasks. For example, she suggested that students can analyze grammatical features of their task performances after transcribing what they said. Long (1991) stated that focus-on-form during TBLT can be achieved through corrective feedback. He has proposed that focus on form must be reactive in nature, and that this reactive focus occurs while a task is being performed. Ellis (2016) explained four ways to incorporate focus on form into task-based language teaching: text-enhancement, corrective feedback, pre-task planning, and task-repetition. Skehan (1998) stated that focus-on-form can also be implemented through pre-task activities. As Ortega (2012) acknowledged, pre-task planning or guided planning can improve learners' accuracy in TBLT. Among all the form-focused instruction, explicit instruction is missing.

Explicit Instruction in TBLT

Explicit instruction refers to attempts to intervene directly in the process of acquisition (Ellis, 2018, p. 112). One primary goal in task-based language teaching is to facilitate incidental language learning as learners convey meaning in their attempts to achieve a task goal. Thus, including explicit instruction in TBLT is controversial (e.g., Long, 2016). On the contrary, some researchers believe that pre-task explicit instruction is beneficial (e.g., DeKeyser, 2003, 2007). Compared to the explicit focus on forms (FonFs) instruction in which a target form is isolated from a meaningful context, form-focus instruction in TBLT is still best characterized as a hybrid type of language instruction in which learners acquire target linguistic features explicitly.

According to Ellis (2018), there are three stages in which explicit instruction can be included: the pre-task stage, main-task stage and post-task stage. The pre-task stage occurs when teachers provide explicit instruction on certain linguistic forms prior to task performance. The target linguistic features are hypothesized to be useful for speakers as they achieve a task goal. In the main-task phase, explicit instruction is usually provided through explicit corrective feedback. For example, Samuda (2001) reported that the participants used the target feature *must* more often when explicit feedback was provided such as *When you are NOT 100% sure about something, you can use must. Not he is business man but he must be a businessman*. In the post-task phase, learners can focus their attention on linguistic form by transcribing their own performance; this task can provide delayed corrective feedback (e.g., Willis, 1996).

Norris and Ortega's (2000) meta-analysis of form-focused studies found that explicit instruction is more effective than implicit instruction, yet few researchers have investigated the effects of a priori explicit instruction on task performance. Research is needed to establish whether explicit instruction followed by a task leads to the acquisition of the target forms and whether explicit instruction impacts how learners perform the task (Ellis, 2018).

Text Enhancement

One way to focus on form in TBLT is for instructors to provide text enhancement (Ellis, 2016). In text enhancement, linguistic features are typographically enhanced through underlining, bolding, or italicizing so that learners notice the target linguistic feature (Doughty, 1991; Lee & Huang, 2008; Sharwood-Smith, 1993). Such enhancements produce a high probability that participants notice the target linguistic forms, as their salience is increased (See Doughty, 1991; Lee & Huan, 2008; Sharwood-Smith, 1993 for studies concerning the effectiveness of typographical enhancements).

Lee and Huang examined 20 text-enhancement studies in their meta-analysis and concluded that text-enhancement has an overall positive effect, but the effect size is quite small. Lower proficiency learners can struggle to engage in processing both comprehending the meaning of the text and consciously attending to linguistic form (Ogawa, 2019). Even if learners notice the target form, they might not acquire it and it might fail to enter long-term memory; therefore, Ellis (2016) recognized the importance of combining text enhancement with other instructional techniques that encourage intentional learning.

Peer Feedback

Another type of form-focused instruction is corrective feedback (CF). Corrective feedback is considered effective in promoting noticing and is thus conducive to L2 acquisition (Lyster & Saito, 2010; Mackey & Goo, 2007; Russell & Spada, 2006; Sato, 2017). However, the main purpose of corrective feedback in these previous studies was for teachers to provide feedback so that the students could acquire accurate grammatical forms. Oral corrective feedback research is primarily concerned with the effect of corrective feedback on learners' acquisition of targeted linguistic forms (Ellis, 2010).

Peer corrective feedback (PCF) might be less effective than teacher corrective feedback because many L2 learners might not feel confident enough to provide corrective feedback or they might trust teacher corrective feedback but not corrective feedback from peers. However, research shows that peer corrective feedback has positive effects on L2 learners' language development (Kim, 2013; Sato & Lyster, 2012; Sippel & Jackson, 2015). The main difference between teacher corrective feedback and peer corrective feedback is that L2 learners can act both as a receiver and provider of feedback; thus, peer corrective feedback provides dual benefits (Sato, 2017). From a feedback providers' point of view, they first need to detect errors in the input their peer produces. In order to do so, feedback providers must notice the gap between the error and the target rule. Noticing provides learners with the opportunity to compare the peer's error and their own interlanguage rule system and notice that they might make the same error and correct it internally. This cognitive process can contribute to restructuring the feedback provider's L2 knowledge (Sato, 2017). On the other hand, from a receiver's point of view, peer corrective feedback can trigger noticing and push the speaker to modify the original utterance. This cognitive process can be explained using Levelt's speech model. First, peer feedback can contribute to noticing the gap between inaccurately proceduralized knowledge and the input because the peer feedback involves checking the accuracy of language production. Second, noticing triggered by monitoring is beneficial for L2 development while both processing input and producing output (Kormos, 2006).

Although the effects of peer corrective feedback are theoretically and empirically acknowledged, L2 learners might find it difficult to correct their classmates' errors due to the social and psychological nature of peer corrective feedback. Philp, Walter, and Basturkmen (2010) explained that L2 learners hesitate to provide peer corrective feedback because they feel less confident of their proficiency (e.g., readiness to correct as a learner) and social relationship (e.g., face saving). Related to this issue, Sato also suggested that it is essential for L2 learners to be trained how to interact with each other and how to provide corrective feedback because providing corrective feedback to peers is influenced by the social dynamics among peers. Therefore, teachers need to create a positive mindset toward peer corrective feedback so that the learners can provide peer feedback and thereby promote L2 development.

To date, few studies have been carried out to explore the role of peer feedback on a speaking task, especially on the use of formulaic language. It is uncertain if and how L2 learners are able to contribute to acquisition through peer feedback and on their usage of target linguistic features. Experimental studies of pedagogical interventions are needed.

Pre-Task Planning

Foster and Skehan (1999) explored the effects of three sources of strategic planning—teacher-led, solitary, and group-based planning—on learners' speaking performances in decision-making tasks. The researchers also examined the effects of planning by investigating form-focused planning and meaning-focused planning. The participants, 63 adult ESL students in the United Kingdom, engaged in a decision-making task in which they had to decide which person should be thrown from a hot air balloon. The researchers divided the participants into six planning conditions. In the no-planning condition (*n* = 12), the students immediately started the task without planning. In the teacher-led language-focused condition (*n* = 11), the teacher explained the structural objective and the use of modals and conditionals using reasons to achieve a task goal through examples (e.g., I take care of hundreds of sick people → If you threw me out, many people might die). In the teacher-led content-focused condition (*n* = 8), the teacher led a discussion concerning ideas that the characters might use to defend their right to stay in the balloon. In the group planning language-focused condition (*n* = 12), the participants discussed the language they could use and confirmed that the English they produced was formed correctly by consulting one another. In the group planning content-focused condition (*n* = 8), the students discussed the reasons for their decisions. In the solitary planning condition (*n* = 12), the students autonomously decided whether they would focus on form or content. A 2 x 2 research design was used to contrast the source of planning (two levels: teacher-led and group based) and focus of planning (two levels: form and meaning).

The findings showed that solitary planning, the teacher-led language-focused condition, and the teacher-led content-focused condition affected CALF measures positively. The teacher-led condition was significantly better than no-planning, solitary planning, and group-based planning in terms of syntactic accuracy. The group planning condition group did not perform well compared to the other groups. The authors concluded that teacher-led planning can help learners enhance syntactic accuracy regardless of whether the focus is on meaning or form because to provide content preparation, correct and complex language is inevitably provided, as a model, even though the focus of the preparation is not the language itself (p. 241). The authors also emphasized that teacher-led planning can enhance syntactic complexity and fluency, a finding that implies that teacher-led planning produces the most well-balanced output among the three sources. In other words, both syntactic complexity and syntactic accuracy benefit from teacher-led planning, a conclusion that is not supported by Skehan's Limited Attention Capacity hypothesis. Although different sources of planning provided different benefits, their study did not clearly differentiate the effects of form-focused or meaning-focused planning. Therefore, as the authors stated, it is necessary to have better research designs to identify the effects of these two types of planning.

Geng and Ferguson (2013) replicated Foster and Skehan's (1999) study by re-examining the effects of participatory structure—solitary planning, pair-work, and teacher-led—on CALF measures. The participants, 32 ESL students studying in the United Kingdom, completed a decision-making task and an information exchange task. In the decision-making task, the participants selected items useful for survival on a desert island. In the information exchange task, they made recommendations to a friend regarding what to see and places to visit in Japan. The participants were divided into solitary, pair-work, teacher-led, and no-planning groups. All of the treatment groups had 10 minutes to think about the content and the forms they might use during the planning time. The participants in the solitary planning group had 10 minutes to plan individually by focusing on both content and form. The participants in the pair-work group also had 10 minutes to discuss both content and form with their partner. In the teacher-led planning condition, the teacher first focused on the content, and then focused on grammar and vocabulary. The researchers calculated complexity by dividing the total number of clauses by the total number of AS-units, accuracy was computed using the number of grammatical errors per 100 words, and fluency was operationalized as words produced per minute excluding false starts and pauses.

Geng and Ferguson found that the no-planning condition was significantly inferior to all of the planning conditions—individual, pair-work, teacher-led planning—for all CALF components. Unlike Foster and Skehan (1999), who found that group-work planning did not benefit students' performances, Geng and Ferguson found that pair-work planning benefitted fluency. However, teacher-led planning did not lead to statistically significant increases in syntactic accuracy, although the descriptive statistics indicated that there was a positive effect. The researchers speculated that the lack of significance was caused by the small sample size and resulting lack of statistical power.

While Geng and Ferguson examined pre-task planning in terms of who engaged in it, Kawauchi (2005) examined three types of solitary planning: writing, rehearsing, and reading. The participants, 39 Japanese university students, first performed a picture narration task without planning. In subsequent weeks, three groups took part in the three planning conditions. In the writing condition, the participants had 10 minutes to write what they wanted to say when they performed the same task used in the no-planning condition. In the rehearsal condition, the participants rehearsed the task for 10 minutes by saying aloud what they had tried to say the first time they did the task. In the reading condition, the participants read a model passage of the task performance silently for 10 minutes and thought about how they could do the task again. Kawauchi analyzed the students' oral performance quantitatively using CALF indices and analyzed the transcripts qualitatively. For the quantitative analyses, Kawauchi assessed fluency using speech rate and the frequency of repetitions. The number of clauses per T-unit, T-unit length, subordinate clauses, and the number of word types were analyzed for syntactic complexity. Correct usage of past tense verbs was analyzed for syntactic accuracy. Qualitative analyses were conducted using transcriptions of the students' oral performances to determine differences in linguistic features in the planning and no planning condition.

There were two main results. First, no statistically significant differences were found among the three planning types on the CALF measures. Second, the analyses of the transcripts revealed qualitative differences. In the reading condition, the participants borrowed lexis and multi-word units from the reading passage; thus, Kawauchi concluded that reading the model passage led to more accurate lexical use, while writing and rehearsing helped the learners generate more ideas.

To test the hypothesis that form-focused instruction improves learners' oral performances, Mochizuki and Ortega (2008) examined whether guided planning with L2 audio input and grammar guidance influenced students' accuracy when providing relative clauses. The participants, 56 EFL Japanese high school students, performed an oral story-retelling task. They were divided into three planning conditions. The students in the no-planning group (*n* = 17) retold the story immediately after listening to the story while looking at pictures. The students in the unguided planning group (*n* = 20) engaged in planning for five minutes after being exposed to the same aural and picture stimuli. The students in the guided planning group (*n* = 19) also engaged in planning for five minutes, but they received a handout explaining how to make sentences with relative clauses. The researchers analyzed the participants' oral performances using the following measures. Syntactic accuracy was operationalized as the number of relative clauses. Syntactic complexity was operationalized as the number of words divided by the number of T-units, the number of subordinate or dependent clauses divided by the number of T-units, and the number of relative clauses divided by the number of T-units. The mean number of words per minute was measured to estimate oral fluency.

The results showed that the guided planning group produced relative clauses significantly more frequently than the no-planning group and the unguided planning group. In addition, the guided planning group used relative clauses more accurately. The researchers interpreted this finding as indicating that instruction focused on linguistic form prior to the performance of a task benefits learners in terms of both frequency of use and the accuracy of the specific form. Two issues qualified the results of the study. First, the authors did not know what the participants in the unguided planning condition did while planning. Second, the cross-sectional design could not show development over time; thus, the authors stated that longitudinal treatments need to be used in the future.

In sum, all researchers have found that teacher-led planning is beneficial in terms of producing gains (e.g., Foster & Skehan, 1999) in syntactic accuracy (e.g., Geng & Ferguson, 2013), and the use of more varied lexis and multi-word units (Kawauchi, 2005), but not to a statistically significant level. While Geng and Ferguson and Kawauchi did not clearly indicate whether they used a meaning-focused or form-focused approach, Foster and Skehan (1999) and Mochizuki and Ortega (2008) explored focus-on-form with teacher-led planning. However, the teacher-led focus on form approach produced mixed results. Mochizuki and Ortega found that form-focused instruction enhanced the syntactic accuracy of learners' oral performance. On the other hand, Foster and Skehan (1999) did not find differences between form-focused and meaning-focused instruction. To understand the reason for the mixed results, researchers need to know (a) what learners do during pre-task planning and (b) to what extent they apply their planning during their performances. In the next section, I present three studies in which these issues have been addressed using post-task interviews and a think-aloud protocol.

Ortega (2005) analyzed interview data after her participants completed a task. She examined how the learners used metacognition when planning for the tasks. The participants were 45 university students studying Spanish in an American university. Ortega used retrospective interviews to elicit the participants' metacognitive perceptions about their strategic tasks planning. Her analysis showed that 59% of the participants said that planning helped them perform better on a storytelling task. For example, the learners used the planning time to organize and formulate thoughts and write notes to help lexical retrieval. On the other hand, 41% of the participants did not think that planning was beneficial because for example, the tasks were too simple to require organization and there was a lack of transfer to on-line performance (e.g., I forgot what I had practiced).

The main findings from Ortega (2005) were as follows. First, the learners focused on both form and content during the planning stage. Second, language proficiency influenced how the learners used the planning time. For example, advanced-proficiency learners attempted to have a well-balanced commitment to retrieval and rehearsal, while low-intermediate learners were likely to focus on retrieval strategies in order to find correct vocabulary. Lastly, she found that the presence of authentic listeners enabled the speakers to speak more comprehensibly in order to meet the listeners' needs. Ortega's qualitative analysis effectively revealed what the learners did when planning.

One limitation of Ortega's qualitative study is that it did not clearly determine the relationship between the learners' strategies during pre-task planning and their oral performance. Pang and Skehan (2014) expanded Ortega's qualitative research in a study of what L2 learners say they do when they plan with how they performed after planning. The participants were 48 university students in Macao who completed a picture descriptive narrative task. After the participants completed the task, retrospective interviews were conducted. Using Levelt's (1989) speech model, the coding scheme that emerged from the retrospective interview was macro planning, micro planning, lexical and grammar planning, and metacognitive planning. Macro planning (e.g., scan then describe or look at each picture) and micro planning (e.g., understand pictures in detail, plan how to tell the story, and organize the ideas developed from pictures) are associated with the conceptualizer in Levelt's speech model because they concern what to say. Lexical and grammatical planning (e.g., lexical retrieval, correct use of verb tenses) was associated with the formulator because speakers search for linguistic forms to express their ideas. Pang and Skehan also added metacognitive codes such as rehearse, memorize, and take notes (e.g., the notes taken are helpful to structure a clear story, rehearse to be fluent, and rehearse to check whether what is planned is logical or clear).

Pang and Skehan found that macro planning had little to do with the students' performance. On the other hand, micro planning was associated with syntactic complexity and fluency; thus, when the students planned small details, they were more likely to pause less and produce more subordination. Lexical codes had both positive and negative associations with all the CALF indices. For example, the lexical code *try to remember the words used in the task* had a positive association with syntactic accuracy, but it had a negative association with fluency as assessed using mid-clause pausing. Interestingly, the grammar codes revealed no associations with syntactic accuracy, and negative relationships with subordination and pausing; thus, when the students thought about grammar, they did not perform well in those two areas.

Sangarun (2005) investigated whether specific pre-task planning foci differentially affect task-based performance. The participants were 40 Thai high school students who performed an instruction task and a monologic argumentative task. In the instruction task, the participants left a message on a telephone answering machine canceling an appointment to meet an interlocutor. In the argumentative task, the participants gave their opinions about school uniforms. The participants were divided into four planning conditions: no-planning (*n* = 10), meaning-focused planning (*n* = 10), form-focused planning (*n* = 10), and both meaning and form focused planning (*n* = 10). The three experimental groups—form, meaning, form/meaning—were given 15 minutes to plan for each task. While they planned, they completed a think-aloud protocol, and after the tasks, they participated in retrospective interviews.

Sangarun analyzed the extent to which the participants applied their pre-task plans. First, she analyzed the think-aloud protocol based on three categories; communicative goal setting, meaning planning, and form planning. She then analyzed the participants' application using the number of planned ideas per T-unit, the number of unplanned ideas per T-unit, the number of planned grammatical structures that appeared in the task speech, and the number of unplanned grammatical structures per T-unit that appeared in the task speech using a 4 x 2 design. The independent variables were the four planning conditions and the two task types. The dependent variables were CALF indices.

Sangarun reported that the think-aloud protocols showed that approximately 80-90% of the students' planning was focused on meaning regardless of the type of pre-task planning. Positive effects were found for the meaning/form planning condition for the instruction task, and for all the planning conditions on the argumentative task. Planning that combined meaning and form seemed to be more beneficial than planning focused on each component separately. Not only did it promote fluency, but it also encouraged the participants to pay more attention to grammatical accuracy. The author explained that this finding occurred because the participants who engaged in form/meaning planning decreased the processing load in the conceptualizer, the formulator, or both; thus, they were able to place more attention on grammatical accuracy during the task.

Park (2010) conducted a study focusing on the process of how the participants utilize planning time in order to examine how pre-task instruction and planning influence learners' focus on form during task-based interaction. The participants, 110 Korean EFL university students who performed two narrative tasks, were divided into four groups: general instruction with planning, general instruction without planning, specific instruction with planning, and specific instruction without planning. For general instruction with planning, the students were provided with 10 minutes of planning time without any specific instructions regarding what to focus on. For general instruction without planning, the participants were not provided with planning time. For specific instruction with planning, the participants were provided with 10 minutes of planning time and were instructed to focus on lexical or morphosyntactic form. For specific instruction without planning, the participants had no planning time, but they were instructed to focus on lexical or morphosyntactic form while performing the tasks.

The speakers' oral performances were transcribed and analyzed to determine how many times they used lexical and grammatical Language-Related Episodes (LRE) during the tasks. LREs occur when student pairs talk about linguistic form, ask questions about linguistic form, and correct their language use. For example, when two students were trying to describe a scene in which four boys cut in line to take a bus ahead of three other boys, one student produced the word *ahead* and another student said, "in front of? no, that's not right." The researcher considered this as a lexical LRE, in which the students talked about lexical choice.

A 2 x 2 x 2 repeated-measures ANOVA was conducted. The independent variables were (a) pre-task instruction (general or specific), (b) planning (no planning or 10-minute planning), and (c) language focus (lexical or morphosyntactic). The dependent variable was the LREs that took place during the tasks. Park found that if learners receive no specific instructions regarding what to plan for, they tend to generate ideas, which leads to improved oral fluency. On the other hand, when the learners were told to focus on morphosyntax, they paid more attention to it compared to the students who had received no specific instructions to do so. Park also noted that learners might decrease their cognitive load before performing the task if they have already planned what to say because this strategy can give learners more working-memory capacity to attend to linguistic form during the tasks.

In summary, pre-task planning generally improves learners' oral performance compared to no planning. Most researchers have reported that teacher-led planning results in well-balanced CALF gains (e.g., Geng & Ferguson, 2013; Kawauchi, 2005; Mochizuki & Ortega, 2008), yet mixed results have been reported for syntactic accuracy. For example, Foster and Skehan (1999) found that teacher-led planning did not help learners improve syntactic accuracy significantly, while Mochizuki and Ortega (2008) found that teacher-led planning did lead to gains in syntactic accuracy. One possible reason is that it is almost impossible for learners to completely separate a focus on form from a focus on meaning because meaning and form are highly interrelated, and grammar exists to enable language users to express different communicative meanings (Nunan, 2004, p. 3). As Sangarun (2005) found, most participants used planning time to focus on meaning, and even if speakers focus on form, they try to retrieve lexis rather than morphosyntax (Ortega, 2005; Park, 2010). Park pointed out that learners usually attend to meaning because it is more useful to think about content due to the communicative nature of tasks. This hypothesis is supported by Levelt's (1989) model of speech production, which states that conceptualization occurs first, then, speakers formulate the language representations. Finally, speakers articulate the message.

One limitation of form-focused planning studies is that no task-based researchers have demonstrated that planning has long-term effects on students' language development. Previous researchers have mainly examined the immediate effects of planning on oral performance, rather than acquisition. Providing 10 minutes of planning time cannot lead learners to acquire native-like target forms because interlanguage development occurs gradually and it requires internal restructuring (Park, 2010, p. 10). Therefore, researchers need to consider how pre-task planning can help learners' language development over time.

Online Planning

According to Yuan and Ellis (2003), online planning refers to the process by which L2 speakers attend carefully to the formulation stage during speech planning and engage in pre-production and post-production monitoring of their speech act (p. 6). Their definition concerns careful online planning, which takes during task performance and at the formulation stage (Levelt, 1989). In the formulation stage, speakers have ample time to plan their speech and make use of the allotted time to carefully attend to their oral performance. Because L2 speakers have ample time while they perform, they are able to attend meaning and form fully, which improve their oral performances.

Online planning process is strongly related to Levelt's speech model. Because L2 learners have a limited processing capacity and cannot attend fully to all aspects of CALF (Skehan, 1998), learners with limited L2 proficiency find it difficult to attend to meaning and form at the same time. This concept leads to the idea that the allotted time on task matters because L2 learners can take as much time as possible to carry out conceptualization, formulation, and articulation. For example, when L2 learners are given a short period of time to complete a task, they might need to conceptualize, formulate the preverbal message, and articulate it as quickly as possible, which might interfere with their attempts to produce grammatically accurate performances. In order to understand L2 learners' task performances in CALF more clearly, Ellis (2018) suggested that more research is needed to investigate the factors of online planning.

Pre-task planning, when it is not guided, allows L2 learners to direct their attention to Levelt's first stage of concepturalization because they think about what to say during pre-task planning; in contrast, online planning allows learners to attend more closely to formulation (Yuan & Ellis, 2003) because even if the learners attempt to think about how to say something in the formulation stage during pre-task planning, it is unlikely that they will remember the pre-planned grammar when they are performing task; thus, they will be obliged to formulate grammar while performing the task. Indeed, on-line planners must attend to the conceptualization stage; therefore, online planning leads to both conceptualization and formulation stage. When learners have the opportunity to engage in careful online planning, syntactic accuracy and syntactic complexity are likely to improve (Ellis, 2018; Yuan & Ellis, 2003; Ahmadian & Tavakoli, 2011) because planning while performing a task potentially allows learners to overcome the trade-off between syntactic accuracy and syntactic complexity (Ellis, 2018). L2 learners can access linguistic knowledge and also carry out the monitoring that aids syntactic accuracy.

To summarize, online planning plays an important role for L2 speakers to monitor in the formulation stage. Researchers have examined the effects of online planning by providing sufficient time on task, yet few researchers have investigated what learners pay attention to and how they monitor their language usage during online planning. More research should be conducted to better understand this issue.

Limited Attentional Capacity Hypothesis

The interrelationships among the CALF components is reviewed in this section. The Limited Attentional Capacity Hypothesis (Skehan, 1998) is based on research in the field of cognitive psychology indicating that learners' working memory and attentional capacity are limited (VanPatten, 1990). Therefore, Skehan has stated that improving performance in one area can come at the expense of performance in other areas. For example, when learners try to speak more fluently, they can experience difficulty paying attention to linguistic form and therefore syntactic accuracy can suffer. A second example is that when learners try to speak more accurately, they might not produce syntactically complex sentences. Skehan (1998) stated that high-level performances for L2 learners can occur in two out of the three CAF components, but not in all three although high-proficiency L2 users can do all three well. According to Skehan's hypothesis, increases in fluency can be accompanied by increases in syntactic accuracy or syntactic complexity, but not both. In other words, these two variables do not increase in tandem.

Given that learners have limited working memory capacity, pre-task planning can be beneficial because it can ease the cognitive pressure on learners' limited working memory capacities, as they can activate both concepts and linguistic forms during pre-task planning. For example, learners can generate ideas about what to talk about during pre-task planning, and this can allow them to pay more attention to linguistic form. Therefore, pre-task planning is hypothesized to influence learners' oral performances positively (e.g., Ortega, 1999; Skehan & Foster, 2005; Yuan & Ellis, 2003). Most previous examinations of pre-task planning support the Limited Attentional Capacity Hypothesis (e.g., Yuan & Ellis, 2003) as well as the idea that oral fluency and syntactic complexity often improve, while syntactic accuracy rarely does so.

Proceduralization and Automaticity

Automatization is a key characteristic of speaking development because once L2 speakers automatize access to and use of linguistic representations, they are able to speak more efficiently and accurately (Segalowitz, 2003). Anderson's Adaptive Control of Thought (ACT) theory (Anderson, 1983, 1987) is a starting point for understanding automaticity. ACT theory hypothesizes that skill acquisition involves a transition from declarative knowledge to procedural knowledge. Declarative knowledge, which is a form of intellectual knowledge, allows learners to describe skill-relevant knowledge. For example, many Japanese learners of English can explicitly explain how to form past tense verbs in English. According to ACT theory, declarative knowledge is the basis for developing procedural knowledge, which is knowledge of how to do something; as such, it involves behavior that people cannot adequately describe. For example, most Japanese

L1 speakers can create accurate grammatical constructions in Japanese without being able to explain the rules of the language. The transition from declarative knowledge to procedural knowledge through the application of production rules is called procedualization. Although rules are explicit initially, the repeated application of the explicit rules in a consistent manner creates rules that are automatic and implicit. This stage of skill acquisition is called automaticity (Segalowitz, 2003).

Towell et al. (1996) argued that the proceduralization of linguistic knowledge is the most important factor in the development of fluency in advanced second language learners. Anderson's (1983) ACT theory shows how procedural knowledge can be developed; thus, Towell et al. argued that the increase in mean length of run is mainly attributable to the procedurallization of different kinds of knowledge, including procedural knowledge of syntax and lexical phrases. Fluency development is shown by an increase in the length and complexity of utterances between pauses, that is, an increase in the mean length of fluent run while pause lengths remain stable.

Segalowitz (2003) explained that automaticity in language learning is characterized by more efficient, more accurate, and more stable performances. Automatization is beneficial for language learners because automatic processing consumes fewer attentional resources compared to controlled processing; thus, speakers can use their attentional resources for other purposes (p. 400) and this development can eventually improve the quality of performance. Automaticity is strongly associated with fluency (De Bot, 1996; Segalowitz, 2003; Towell et al., 1996) because the automatic execution of L2 speaking performances such as lexical retrieval, pronunciation, and grammatical processing will promote fluency; therefore, it is important to enhance automaticity in second language classrooms (De Bot, 1996; Segalowitz, 2003).

Logan's Instance Theory

In contrast to the positions of Anderson's rule-based automatization theory, Logan (1988) views automatization as an instance-based theory in which L2 learners' retrieve instances from memory. Logan's instance theory hypothesizes that novices use algorithms when completing tasks with which they have little processing experience. This notion applies to L2 learners' attempts to construct sentences using grammatical rules. When the learners have produced a sentence (i.e., an instance), they can eventually perform the same task using memory retrieval of the solution. In this theory, high-proficiency speakers retrieve language representations from memory and as these memory-based processes become stronger, they replace their use of rules and algorithms with instances. For example, in lexical decision tasks, reaction times decrease for specifically practiced words, but not for new words, which explain that experiences with previous instances lead to automatization in language performance.

Logan's instance theory explains why native speakers are fluent; many of their utterances are formulaic sequences that they can automatically retrieve. In Logan's theory, similar to Anderson, learners are not able to automatically retrieve a chunk unless they have extensive practice. Kormos (2006) pointed out that there are problems with both Anderson and Logan's models in L2 learning. Anderson does not explain why speech is so formulaic, and Logan does not explain why speakers can generalize chunks of language to new contexts.

To summarize, Anderson's ACT theory, which concerns the transition from declarative knowledge to proceduralized knowledge, is called rule-based because the rules that L2 learners acquire initially are proceduralized by applying the same rules repeatedly. Logan's instance theory, which concerns the memory retrieval, is called instance-based because learners can retrieve appropriate, previously encountered instances from memory.

Task Repetition

Anderson's ACT theory has been applied to second language learning and it plays an important role in second language acquisition research. This discussion raises the question of how teachers can promote proceduralization and automaticity: Although task repetition promotes procedualization (e.g., De Jong & Perfetti, 2011), merely repeating the same rules has been criticized because it does not provide a meaningful context in which students genuinely need to communicate. Automaticity is best achieved by the repeated use of language rules in a context of authentic communication (DeKeyser, 2003; De Ridder, Vangehuchten & Gómez, 2007; Segalowitz, 2003).

Task repetition is helpful for developing fluency because repetition allows learners to activate concepts and linguistic forms so that they are more easily and quickly accessed. Based on the limited attentional model of speech production (Skehan, 1998), low-proficiency L2 learners face a number of challenges in the speaking process from conceptualization to articulation because of the demands of thinking of a preverbal message and formulating the message efficiently. Repetition might reduce the attentional demands on learners as they conceptualize, encode, and monitor their messages.

One effective activity for developing oral fluency is the 4/3/2 task (e.g., Boers, 2014; Nation, 1989; Nation & Newton, 2009; Thai & Boers, 2016). In this task, students talk about the same topic for 4 minutes, then 3 minutes, and finally 2 minutes. When the students talk about the same topic three times with increasing time pressure to perform more quickly, they must speak faster. This activity can be used in language classrooms to foster speaking fluency.

According to Nation (1989), the 4/3/2 task has three important features: repetition, a reduction in time, and a change of audience. These features directly affect fluency by encouraging L2 speakers to focus on the meaning under a time constraint. Nation also stated that repetition can have an effect on grammatical accuracy (p. 380). One reason is that repetition can result in the provision of more time for monitoring, which allows speakers to reduce grammatical errors. Another reason is that repetition can have a local rather than a general effect, which allow L2 speakers to monitor their speech.

De Jong and Perfetti (2011) investigated whether the 4/3/2 task would lead to a long-term increase in oral fluency with 24 adult ESL learners in the United States. The participants were randomly assigned into repetition, no-repetition, and control groups. All participants completed speaking tests 2 to 3 days before the training started (Time 1), immediately after the training (Time 2), and 3 weeks after the training (Time 3). The tests were a 2-minute personal-story monologic task on different topics. Two or three days after Test I, the 4/3/2 condition was implemented for two groups. The control group (*n* = 5) did not engage in the 4/3/2 task, the repetition group (*n* = 10), spoke on one topic three times (4 minutes, 3 minutes, and 2 minutes), and the no-repetition group (*n* = 9) spoke on three different topics for 4, 3, and 2 minutes. The participants talked about personal topics such as *What do you think about pets?* and *Who is your favorite artist?*. After the three-week treatment, the participants completed the Time 2 performance. After Time 2, the control group started the repetition condition for three weeks. During that period, the repetition group and no-repetition group did not do the 4/3/2 task, so the researchers could determine whether they retained fluency gains on the delayed posttest performance. After three weeks, the participants took the Time 3 delayed posttest.

The researchers analyzed four fluency measures: mean length of fluent runs in syllables, mean length of pauses in seconds, phonation/time ratio and articulation rate in syllables per minute. They also took a preliminary look at the role of word repetition during the 4/3/2 training session. They counted how many words overlapped across the three deliveries and they categorized the repeated words into topic-related and non-topic related words. Examples of topic-related words were *Beckham, soccer, sport*, and *play* when the topic was about sports. Examples of non-topic related words were *favorite, know, like*, and *make*.

First, the results showed that fluency improvements during the 4/3/2 task were the result of proceduralization by the participants who repeated the same topic. Second, the fluency improvements were retained over four weeks and they transferred to new topics for those in the repeating groups. This result was plausible, as students should proceduralize and eventually automatize the linguistic features they use because of the time pressure and task repetition inherent in the task design (De Jong & Perfetti, 2011, p. 538). Third, the students in the repetition group repeated more words than those in the no-repetition group. The students who repeated the same words across three deliveries showed great improvement at Time 2.

This study had two limitations. First, De Jong and Perfetti only measured fluency; however, because oral performance is multi-dimensional, other components, such as complexity and accuracy, need to be analyzed, as this would result in a better understanding of the interrelationship of the CALF variables through the repetition of the oral speaking tasks. Based on Skehan's Attentional Capacity Model, learners might display trade-off effects if they attend primarily to one component. Therefore, examining all three components is crucial in speaking research. Second, as De Jong and Perfetti recognized, it was not clear what kind of linguistic knowledge was proceduralized due to the treatment. Individual words and multi-word units were repeated relatively few times. Further investigation of specific linguistic features needs to be conducted.

To investigate whether a shrinking time condition can promote syntactic accuracy, Boers (2014) compared learners' performance under a time-shrinking condition and a time-constant condition. Previous researchers have reported that task repetition can lead to greater syntactic accuracy (Fukuta, 2016; Gass, Mackey, Alvarez-Torres & Fernández-García, 1999), lexical sophistication (Gass et al., 1999), and lexical variety (Fukuta, 2016). However, the participants in these previous studies were provided a repetition opportunity with the same time on task. Boers' investigated whether shrinking time in a 4/3/2 task helped learners improve syntactic complexity and syntactic accuracy in addition to oral fluency. He investigated the importance of shrinking time because the need to speak quickly can compromise linguistic accuracy according to Skehan's Limited Attentional Theory. The participants were 10 ESL adult ESL learners in New Zealand who were asked to select two topics they felt comfortable talking about. To counterbalance the task order, five participants did the 4/3/2 task first, while the other five participants did the 3/3/3 activity first. The mean quantitative changes in the CALF indices between the first delivery and the third delivery are summed up for the two task conditions. Boers compared the mean changes on CALF indices between the first and third deliveries. The results showed that the learners improved fluency in the shrinking time condition, as was found by De Jong and Perfetti (2011). There were no significant changes in syntactic complexity and lexical sophistication in either condition. On the other hand, there was a moderate negative correlation (r = -.56; p = .01) between the learners' speech rate gains and improvements in syntactic accuracy.

Thai and Boers (2016) conducted a similar study in which they examined the 4/3/2 speaking task with and without time pressure. This study was similar to Boers (2014); however, Thai and Boers used a between-subjects design. The participants were 20 tenth grade EFL students in Vietnam who talked about the same topic, *favorite movie*. Ten students were in the 3/2/1 condition, while the other ten students were in the 2/2/2 condition. The researchers analyzed all 60 speeches (20 participants x 3 deliveries) using CALF indices. The results indicated that oral fluency (syllables per minute) improved significantly under the time-shrinking condition (3/2/1), but there was no significant development under the 2/2/2 condition. There was no significant improvement in syntactic complexity in the time-shrinking condition, but there was a significant gain in syntactic complexity (mean ratio of clauses per AS-unit) in the time-constant condition. The researchers also analyzed lexical diversity (type-token ratio) and lexical sophistication (usage of difficult words), but there was no evidence of improvements in lexical sophistication in either condition. There was also no improvement in syntactic accuracy (error free AS-units) in the time-shrinking condition, while there was a significant improvement in the time-constant condition. Finally, a negative correlation was found between oral fluency and syntactic accuracy (*r* = - .41; *p* = .037) and oral fluency and syntactic complexity (*r* = - .41; *p* = .036), findings that supported Skehan's Limited Attentional Hypothesis (p. 229).

The two studies conducted by Boers (2014) and Thai and Boers (2016) indicated that decreasing time is beneficial for fluent performances but not for producing more syntactically complex and accurate performances. As Thai and Boers acknowledged, the findings might differ if a longitudinal study had been conducted. Therefore, investigating the extent to which the 4/3/2 task helps learners improve CALF measures longitudinally is essential to understanding foreign language learners' language development. If learners gradually improve fluency through engaging in the 4/3/2 task, they might be able to pay more attention to linguistic form. In addition, Thai and Boers (2016) also reported that there were no changes in lexical sophistication between the first and the latter delivery. However, providing learners with model input and encouragement to use the input in their own speech might result in greater lexical sophistication (Boers, 2014, p. 231).

In sum, as shown in previous studies (e.g., Boers, 2014; De Jong & Prefetti, 2011, Thai & Boers, 2016), the 4/3/2 task improves oral fluency. In the 4/3/2 task, learners are familiar with the content of what they will say in the first trial, so they can redistribute their focus from conceptualization to formulation in the subsequent performances. Because of repetition and time pressure, automatization and proceduralization of knowledge occurs. Although some previous researchers (e.g., Boers, 2014; De Jong & Prefetti, 2011) have suggested using pre-task planning prior to 4/3/2 tasks, no researchers have investigated the longitudinal effects of pre-task planning prior to the 4/3/2 task.

Transfer Appropriate Processing (TAP)

The Transfer Appropriate Processing (TAP) Theory is rooted in cognitive psychology in the framework of information processing theory. A basic assumption of TAP is that the human mind has limited information processing capacity and that there are constraints on the amount of information learners can pay attention to (Spada, Jessop, Tomita, Suzuki, & Valeo, 2014). Therefore, learners are more likely to remember something similar to what they learned if the cognitive processes activated during learning are the same as those activated during retrieval (Lightbown, 2007; Morris, Bransford, & Franks, 1977; Spada et al., 2014).

Morris et al. (1977) demonstrated that TAP had more explanatory power than the levels of processing approach, in which deeper processing at the time of learning results in more learning than shallower processing. The level of processing approach means that deep processing such as words processed at a semantic level are better remembered than words processed at a shallower orthographic or phonetic level. Morris et al. asked the participants to identify new words to make either semantic or phonetic judgement about sentences and words they heard in the learning condition. They found that those who had experienced orienting questions that drew their attention to rhyme (e.g., Does *BLANK* rhyme with legal?... EAGLE) were more successful in the phonetics retrieval test than those who had been oriented to meaning (e.g., Does *BLANK* had a silber engine…TRAIN). Morris et al. showed that the more important issue is the match between type of learning and the type of test used. When learners try to recall the linguistic item, they also recall aspects of the learning process. Therefore, when there is greater similarity between the learning processes and knowledge retrieval, the chances of successful recall are greater.

TAP perspectives on memory retrieval provides an important message for understanding oral fluency in skilled performance (Segalowitz, 2010) because according to the TAP principle, the fluency and accuracy of memory retrieval depends to a significant extent on the similarity between the cognitive processes that were active at the time of learning and the time when the memory was retrieved. This finding suggests that fluency-promoting learning conditions for L2 learners should elicit cognitive and perceptual processes that are appropriate for transfer to the situations that will be encountered when the learned items are to be retrieved (Segalowiz, 2010).

TAP theory also provides a theoretical justification for including form-focused instruction in the task performance (Ellis, 2018). Learning demands in the classroom often differ from communicative demands in the real-world (Segalowitz, 2010). There is a mismatch between classroom drills and real-world situations because drilling isolated from communicatively meaningful contexts will not help students become used to the patterns present in authentic communication situations. Therefore, TAP helps to explain why L2 learners do not always mobilize the knowledge they have learned in certain classroom activities in the real-world (Larsen-Freeman, 2013; Lightbown, 2007). Form-focused instruction in a meaningful context in the language classroom is beneficial based on the TAP theory.

Formulaic Language

According to Wray (2008), formulaic language refers to large units of processing; in other words, lexical units that are longer than one word. Some formulaic language is sequences of words whose meaning is not entirely predictable from the individual words (e.g., *kick the bucket*). The words in a formulaic sequence are glued together and stored as a single word (Ellis, 1996).

The study of formulaic language is related to automaticity and it also plays an important role in second language acquisition. Learners' automatic access to prefabricated chunks, which are stored in memory, can lead to fluency development (Boers & Lindstromberg, 2012; Segalowitz, 2003; Wood, 2009, 2015) through the repeated use of formulaic language. Linguistic chunks can become part of production rules and retrieved directly from declarative memory without the need for computations in working memory (Wood, 2010, p. 3). If learners process formulaic language automatically, they can use more attentional resources for other areas. Nonetheless, Segalowitz suggested that more research should be conducted to clarify the relationship between automaticity and formulaic language.

According to Boers, Eyckmans, Kappel, Stengers, and Demecheleer (2006), there are three reasons why acquiring formulaic language is believed to be beneficial to L2 learners. First, the mastery of the idiomatic aspects of natural language can help learners sound more native-like. Second, the mastery of formulaic language can help learners to speak more fluently because prefabricated sequences or ready-made chunks can be retrieved faster than sentences can be generated word by word under real-time conditions.

Third, formulaic language can help learners produce more accurate language provided that the pre-fabricated chunks are stored correctly in memory.

Formulaic sequences are grouped into textual, interpersonal, and ideational metafunctions (Butler, 2003; Qi & Ding, 2011). Textual formulaic sequences occur when speakers relate a clause to the preceding text using phrases such as *because of* and *as a result*. Interpersonal formulaic sequences occur when speakers express their opinions using phrases such as *in my opinion* and *kind of*. Ideational formulaic sequences refer to physical or abstract entities that represent patterns of experiences such as *read aloud*, *video games,* or *on campus* (Qi & Ding, 2011, p. 166).

Qi and Ding (2011) examined the extent to which 56 Chinese university students developed their use of formulaic sequences over three years by measuring the frequency, accuracy, and variation in the formulaic sequences they produced. The researchers compared the students' speaking monologues in Year 1 and Year 4. They also compared the students' use of formulaic sequences with those of 15 American college students. They found that the Chinese students overused textual formulaic sequence and underused interpersonal formulaic sequences, possibly because they had read a great deal of English and spoken relatively little. The results also showed that the Year 4 students did not repeat the same formulaic sequences as much as they did when they were in Year 1. The authors concluded that as the students became more advanced, they were able to use a greater variety of formulaic sequences more frequently. Although Qi and Ding examined the frequency of learners' usage and the emergence of formulaic language over time, it was not clear what led them to acquire these formulaic sequences over the three years.

Wray (2000) stated that L2 learners have difficulty acquiring formulaic language for several reasons. First, L2 learners are not exposed to formulaic language frequently. Formulaic language is often omitted in interactions with L2 speakers although formulaic language is commonly used among native speakers of a language. The second reason is that formulaic language is often not taught in foreign language classrooms.

Gatbonton and Segalowiz (1988) made a significant contribution to classroom instruction in terms of a way to automatize formulaic language. They claimed that the traditional way of repeating the target form in monotonous drills is not sufficient because the task lacks a communicative context. Their suggestions demonstrate a way to incorporate the repetition and rehearsal of formulaic language into a communicative task. They called this process *creative automatization* because the students themselves generate and create appropriate utterances based on their understanding of the communicative situation. Automatizing utterances requires students to repeat utterances that occur naturally in normal communicative situations. Gatbonton and Segalowiz suggested that the tasks should be designed to elicit short and memorizable utterances. The formulaic language should be multi-situational so that it is usable in many situations with little or no modification. The rehearsal of short, memorizable formulaic language in communicative contexts promotes fluency with that language.

Wood (2009) conducted a case study of the classroom teaching of formulaic language and fluency development with a female Japanese learner of English. This study took place in an intensive study abroad class, but only the female participant's speaking progress was analyzed. Therefore, Wood described the classroom activity that all the students in the program completed, but he analyzed the degree to which the participant developed speaking fluency and formulaic language through a fluency workshop that consisted of nine hours of instruction over six weeks. The sessions included an (a) input stage, (b) automatization stage, (c) practice and production stage, and (d) free talk stage. In the input stage, the student listened to native English speakers' personal stories. The instructor drew attention to formulaic language and commented on the linguistic and discourse functions of the formulaic language. In the automatization stage, the student shadowed the recorded model at least eight times. The student then did a dictogloss activity by listening to the sentences that included the formulaic language taken from the input passage. In another activity, the mingle jigsaw, the student was provided with slips of paper with key formulaic language from the original text and instructed to memorize them. She was instructed to mingle and share the assigned formulaic sequence with other classmates and to listen to the others' assigned formulaic language until all the students had a chance to record every formulaic sequence. In the practice and production stage, the student did the 4/3/2 task, in which she told personal narratives. In the free talk stage, students in small groups took turns listening to individuals speaking spontaneously about the topics they had been assigned. The participant commented on her speech and reflected on the speed and hesitations in her own speech.

The female participant did a monologue narrative recording before she started the formulaic language instruction and then after the six-week training session. Her speech data were analyzed for speech rate (the number of syllables per minute) and mean length of runs (the total number of syllables divided by the number of runs). The results showed that she made a 13.8% gain (123.2 to 140.2 syllables per minute) for speech rate and a 26.3% (5.1 to 6.4 syllables per run) gain for mean length of runs between the pretest and posttest. Not only did the workshop improve her fluency, it also helped her to improve complexity, as she used a greater variety of formulaic language that native English speakers would use. Before the fluency workshop, she produced 18 tokens of formulaic language, two of which were present in the native English speaker model. After the workshop, she used 52 tokens of formulaic language, 18 of which were present in the native English speaker models. Examples included *when I was a little girl*, *it took about ten minutes*, *in the daytime/nighttime*, and *still now* in the posttest. This result showed that the fluency workshop provided the participant with samples of formulaic language that she added to her repertoire; thus, her utterances became more fluent.

According to Boers and Lindstromberg (2012), ways to foster the use of a greater breadth of formulaic sequences have been examined in many intervention studies, but few researchers have examined the proceduralization of formulaic sequences. Given that formulaic sequences were more proceduralized when memorized than when analyzed syntactically (Yu, 2009), it is worth investigating the degree of development of formulaic language in communicative oral tasks.

To summarize, formulaic language can help L2 speakers sound more fluent because automatic retrieval of pre-fabricated chunks is faster than retrieving phrases word by word. Teaching formulaic language can help learners improve speaking performances (Wood, 2009), yet it is difficult for L2 learners to acquire formulaic language partly because they have little exposure to it, even though it is commonly used among native speakers (Wray, 2000). Indeed, few researchers have examined the effects of teaching formulaic language on L2 learners' speaking development.

Gaps in the Literature

This study is designed to address four gaps in the literature. First, few researchers have explored the longitudinal development of CALF measures through form-focused pre-task planning. Pre-task planning was considered effective, especially in terms of developing oral fluency and syntactic complexity because learners usually attend to meaning while engaged in pre-task planning (e.g., Foster & Skehan, 1999; Geng & Ferguson, 2013; Kawauchi, 2005; Mochizuki & Ortega, 2008 Ortega, 1999). Although learners often have higher fluency and produce more complex utterances after pre-task planning, syntactic accuracy typically does not improve.

When researchers have discussed the limited attentional hypothesis, they have looked at learners' immediate speaking performances rather than speaking development because pre-task planning has usually been examined in cross-sectional designs (e.g., Foster & Skehan, 1999; Mochizuki & Ortega, 2008). Few studies have been conducted to investigate language performance development (Vercellotti, 2017). Considering that language development occurs gradually over time, researchers need to investigate how learners acquire linguistic form rather than merely looking at how learners perform after only one treatment.

Second, L2 learners' proceduralization and language development through the 3/2/1 task is not clear. De Jong and Perfetti (2011) found that repeating the same topics enhanced learners' fluency compared to talking about the different topics in their 4/3/2 task because of proceduralization; however, learners' development in terms of syntactic complexity, syntactic accuracy, and lexical diversity were unknown. Other researchers examined components such as syntactic accuracy and syntactic complexity in addition to oral fluency (Boers, 2014; Thai & Boers, 2016). Thai and Boers and Boers found that a shrinking time condition did not lead to improvements in syntactic accuracy; instead, the speakers' oral fluency improved. Their findings suggested that having the same amount of time is more effective at leading to improvements in syntactic accuracy compared to the shrinking time condition. However, they employed a cross-sectional research design. As Thai and Boers acknowledged, it is essential to investigate to what extent learners improve their performances through the 3/2/1 task longitudinally.

Third, few researchers have investigated communicative adequacy in learners' task performances. Although CALF measurements do not indicate the extent to which learners achieve communicative goals (Pallotti, 2009), they are the primary ways that learners' speaking performances have been analyzed in task-based research. The exclusive use of CALF measures is not sufficient to obtain a valid estimate of successful task performance (De Jong, Steinel, Florijin, Schoonen, Hulstijn, Cresswell & Plano Clark, 2012; Pallotti, 2009; Révész et al., 2016) because even if learners produce complex, accurate, and fluent utterances, they do not guarantee that they have effectively accomplished the task goal. For instance, speakers with greater CALF measures might be off topic or producing poorly organized discourse. Therefore, in addition to CALF measures, human ratings are needed to evaluate to what extent learners successfully achieve task goals.

Finally, few researchers have investigated what aspects of speaking learners prioritize when performing speaking tasks such as 3/2/1. First, it is not clear how learners plan during pre-task planning time because in many previous studies, the pre-task planning time was not controlled (e.g., Orgeta, 1999; Wigglesworth, 1997). As a result, there is no clear understanding of what participants do when pre-task planning and online planning (e.g., Foster & Skehan, 1999: Mochizuki & Ortega, 2008). Second, as noted above, most previous researchers have analyzed students' oral performance quantitatively based on CALF measures (e.g., Boers, 2014; De Jong et al., 2012; Geng & Ferguson, 2013; Ortega, 1999). Although some researchers have qualitatively analyzed think-aloud protocols (e.g., Sangarun, 2005) and retrospective interviews (Ortega, 2005), more studies are needed to investigate learners' task performances using mixed-method research designs. Mixed-methods designs can help researchers understand to what extent learners apply form-focus interventions when performing communicative tasks. Learning more about students' perceptions and strategies during their task performance is an important addition to the quantitative results.

Purposes of the Study

In this study, I examine the effects of two pedagogic interventions—teacher-led planning as input enhancement with target form and a peer-check activity used to pressure to learners to use the target form—to help the participants proceduralize formulaic language through the repetition of communicative tasks.

The first purpose of this study is to explore the longitudinal effects of form-focused pre-task planning. This study is one of the first attempts to examine the effects of pre-task planning longitudinally. The results of many previous studies have supported Skehan's Limited Attentional Hypothesis because one CALF variable improves while one or more others decline. Pre-task planning helps learners to produce more fluent performances because they typically attend to meaning during pre-task planning time.

Skehan (1998) has argued that speakers cannot attend to all three components simultaneously; however, the degree to which trade-off effects apply longitudinally is unclear. Although TBLT is generally a meaning-oriented approach, it should also promote the acquisition of linguistic form because syntactic complexity, syntactic accuracy, and oral fluency should be developed in a well-balanced way. Examining the effectiveness of form-focused pre-task planning can lead to a better understanding of how form-focused instruction can be incorporated into TBLT.

The second purpose is to explore proceduralization through the 3/2/1 task over one academic semester. The repetition of the 3/2/1 task can help students proceduralize and eventually automatize their declarative knowledge (e.g., De Jong & Perfetti, 2011). Automaticity is a vital part of language learning because learners can produce more efficient, more accurate, and more stable performances (Segalowitz, 2003). Acquiring formulaic language is an effective way to improve oral fluency (Boers & Lindstromberg 2012; Gatbonton & Segalowitz, 2005; Segalowitz, 2010; Wray, 2002). Target formulaic language can be proceduralized through repetition in the 3/2/1 task. By investigating the participants' longitudinal development of CALF measures through the 3/2/1 task, researchers can better understand the benefits of task repetition.

The third purpose is to investigate the relationship between communicative adequacy and CALF measures in the 3/2/1 task. It is necessary to investigate the degree to which changes in CALF indices are accompanied by improvements in communicative adequacy, as these are considered to be potentially independent aspects of speaking proficiency (Pallotti, 2009). Many SLA researchers have used analytical CALF measures such as counting mean length of run, subordination, and error-free clauses (e.g., Kawauchi, 2005; Ortega, 1999; Skehan, 1996), while testing researchers (e.g., Elder & Iwashita, 2005) have used analytical ratings in which human judges rate each CALF component. By examining the relationship between CALF measures and human ratings of communicative achievement, researchers can better understand how to interpret CALF measures and their relationship with communicative adequacy.

The last purpose is to investigate what L2 learners prioritize during monologue task performances. Qualitatively examining students' strategies to achieve a task goal efficiently is crucial to understanding how they engage in communicative tasks. The exclusive usage of CALF measurements does not provide enough information about how learners utilize what they have learned from the form-focused instruction. Qualitative findings can contribute to a better understanding of the quantitative results.

Research Questions

The following research questions are investigated in this study.

1. To what extent do the comparison group, teacher-led planning group, and teacher and peer treatment group improve on complexity, accuracy, lexis, and fluency during the 13-week treatment?
2. Do the teacher-led planning group and the teacher and peer treatment group significantly outperform the comparison group in terms of syntactic complexity, syntactic accuracy, lexis, and oral fluency?
3. To what extent do the teacher-led planning group and the teacher and peer treatment groups develop their use of target formulaic language in terms of frequency and variety during the 13-week treatment?

4. To what extent do the comparison group, the teacher-led planning group, and the teacher and peer treatment group develop communicative adequacy during the 13-week treatment?
5. What is the relationship between the analytical measures CALF indices (syntactic complexity, lexical complexity, syntactic accuracy, fluency) and the human ratings of communicative adequacy?
6. What do learners prioritize while performing the monologue speaking tasks?

CHAPTER 3

METHODS

In this chapter, I explain the methods used to answer the research questions in this study. I describe the participants, The English Language Curriculum and the Discussion Course, instrumentation, mixed method research design, procedures, data coding procedures, analysis, Rasch analysis used to answer the research questions.

Participants

The participants, 48 first-year Japanese university students attending a private Japanese university in eastern Japan, were selected based on their enrollment in my seven English discussion classes, which was a required course for first-year students. All participants were informed of the purpose of the study and they signed the Japanese version of the consent form (Appendix A). See Appendix B for the English translation.

Originally, 56 students were enrolled in the classes. Eight students were absent when one of the 3/2/1 recordings was administered, so they were omitted from the analysis; thus, 48 students' data were analyzed in this study. There were 18 male students and 30 female students, with an average age of 18.08 (*SD* = 0.27). Prior to entering the university, the participants completed six years of English classes in junior and senior high school. Thirty-three students took a written English test to enter the university, and eight students entered the university with the recommendation system. These students had high grades in high school or were skilled in areas such as sports. Five students entered the university from the high schools attached to the university. Most of the students studied English in secondary school in order to pass competitive entrance examinations. The participants did not have extended experience speaking English prior to beginning their university studies.

Prior to the first semester, the students took a TOEIC test for placement purposes. Based on their TOEIC scores, they were placed into one of four levels of required English classes (Levels 1, 2, 3, and 4): Students with TOEIC scores 680 and above were placed in Level 1, 480-679 in Level 2, 280-479 in Level 3, and 279 and below in Level 4. None of the participants were in Level 1 or Level 4. Thirty-four participants were in Level 2 and 14 participants were in Level 3. The mean TOEIC score of the 48 participants was 491.15 (*SD* = 48.48). Table 1 shows the information about the participants in each group based on their responses to the Background Questionnaire (Appendix C). See Appendix D for the English version of the questionnaire.

The English Language Curriculum and the Discussion Course

This study was conducted in an English discussion program at a private Japanese university that was founded in 1874. Approximately 20,000 students were enrolled in 10 departments in this university when the study was conducted. All first-year students were required to take a 90-minute English discussion course in their first year of study. The students met weekly for 14 weeks in the spring semester (April-July) and 14 weeks in the fall semester (September-January). After the first semester, the class composition was changed so that the students had different classmates and a different teacher. The data were collected from the spring semester discussion course.

Table 1. *The Participants' Information from the Background Questionnaire*

	Control group (*n* = 12)	Teacher-led group (*n* = 13)	Teacher and Peer group (*n* = 21)
Mean age	18.2 years	18 years	18.2 years
Gender	• Male = 3 • Female = 9	• Male = 7 • Female = 7	• Male = 8 • Female = 13
Major	• Contemporary psychology = 7 • Sociology = 6	• Business = 6 • Literature = 7	• Community and human service = 8 • Sociology = 6 • Literature = 7
TOEIC score	*M* = 519.17 *SD* = 30.66	*M* = 479.29 *SD* = 38.17	*M* = 480.00 *SD* = (57.20)
Method of entry into the university	• Entrance exam = 6 • Attached high school = 3 • Recommendation = 3	• Entrance exam = 10 • Attached high school = 0 • Recommendation = 3	• Entrance exam = 17 • Attached high school = 2 • Recommendation = 2
English learning experience before junior high school	• English conversation school = 4	• English cram school = 1	• English conversation school = 6 • English cram school = 3
English club experience	1	0	1
Length of English learning experiences	6.5 years	6.1 years	7.6 years
Family or friends who you communicate in English	Send messages to friends overseas = 2	0	English speaking father who lived away and communicated twice a year = 1
Oversea experiences	• Age 6-8 in the United States = 1 • Age 3– in the Philippines = 1	0	• Age 0-5 in New Zealand = 1
English language certificate	*Eiken* grade 2 = 5	*Eiken* grade 2 = 6	*Eiken* grade 2 = 7

One important feature of the English discussion course is that the classes were held only in English, as a 100% English policy was employed. The main objective of the English discussion course was for the students to learn how to participate in discussions more effectively and develop speaking fluency. The students discussed various topics such as media, culture, and the environment, and they learned to use formulaic language during extended group discussions.

The English discussion course employed a student-centered approach with a class size of seven to nine students. The small class size maximized student-to-student interaction. The program administrators suggested that the ideal amount of student interaction time was more than 50 minutes in each 90-minute class. The typical lesson plan was as follows. Prior to the class, the students read an 800-1,000 word reading passage about that day's topic. The students then took a three-minute quiz on the reading passage consisting of eight multiple-choice questions. Following the quiz, the 3/2/1 task was conducted.

After the 3/2/1 task had been completed, the students studied the target formulaic language for approximately 15-25 minutes (See Appendix E for the full list of formulaic language). The students completed a short communicative task that required them to use the lesson's target formulaic language. The students asked and answered questions such as *Is it more interesting to talk to friends or family?*, *Is it fun to spend time alone?* and *Is it fun to spend time with family?* using the lesson's target formulaic language.

Following the formulaic language practice, the students usually prepared and practiced in pairs prior to the group discussion. For example, for the topic *friendship*, the students individually selected items from a list of important things for a good friendship (e.g., communicating every day, always telling the truth) and discussed those ideas with a partner for five minutes. These partners were then separated into different groups of three or four members and they engaged in the first group discussion for 10 minutes by discussing two questions *What is important for a good friendship?* and *Which is better, having a lot of friends or one best friend?* After the students finished the 10-minute discussion, they received self-, peer-, or teacher-fronted feedback. During the self-feedback, the students reflected on whether they did a good job using communication skills (e.g., reacting, asking follow-up questions, and agreeing or disagreeing), formulaic language, speaking only English during the group discussion, and thinking about what needs to be improved for the next discussion. During the peer-feedback, the students discussed their use of the above skills. During teacher-fronted feedback, the students listened to the teacher about what types of formulaic language or communication skills were used during the discussion and what needs to be improved for the next discussion. The students then discussed a different question such as *What are good ways to make new friends?* in a different group. First, the students chose three ways to make new friends (e.g., joining a club activity, finding people on social network sites) from the list of choices. They then discussed the topic in pairs; the partner was different from the first partner. These partners were then separated into different groups to form a discussion group in which they conducted the second extended group discussion for 16 minutes.

As part of the English discussion course curriculum, the students took group discussion assessment tests three times a semester in Weeks 5, 9, and 13. The data gathered from the group discussion tests were not a part of the current study. The discussion tests started 45 minutes after the class began on the discussion test days. To ensure that students could practice for the discussion tests, the 3/2/1 task was shortened to a 2/1 task on the discussion test days.

The students' 3/2/1 task performances were recorded in Weeks 2, 8, and 14. The recorded date were later analyzed. The detailed information is provided in the following section. Table 2 shows the discussion course lesson plan for the regular lessons, group discussion test lessons, and 3/2/1 task recording lessons.

Table 2. *Discussion Course Lesson Plan*

Regular weekly lesson plan (Weeks 3, 4, 6, 7, 10, 11, 12)	Group discussion-test lesson plan (Weeks 5, 9, 13)	3/2/1 task recording lesson plan (Weeks 2, 8, 14)
• Quiz (3 minutes)	• Quiz (3 minutes)	• Quiz (3 minutes)
• 3/2/1 task (Treatment phase) (20 minutes)	• 2/1 task (Treatment phase) (15 minutes)	• Practice new (or review) formulaic languages (15 minutes)
• Practice new (or review) formulaic languages (15 minutes)	• Discussion practice (25 minutes)	• Discussion 1 including practice (20 minutes)
• Discussion 1 including practice (20 minutes)	• Group Discussion Test 45 minutes)	• Discussion 2 including practice (25 minutes)
• Discussion 2 including practice (30 minutes)		• 3/2/1 Task Recording (20 minutes)
		• 3/2/1 Recording retrospective questionnaire (5 minutes)

Instrumentation

Background Questionnaire

In Week 2, after the participants finished recording the 3/2/1 task, they completed the background questionnaire (Appendix C) in Japanese. The English version of the background questionnaire is in Appendix D. The purpose of the background questionnaire was to gather information about the participants' educational background and English language proficiency.

3/2/1 Recording Retrospective Questionnaire

The participants answered the Japanese version of the 3/2/1 recording retrospective questionnaire about their performance on the speaking tests immediately after completing their recordings in Weeks 2, 8, and 14 (Appendix F). The English version is in Appendix G. The purpose of this questionnaire was to determine how well the participants thought they did on the tests and how they organized their speeches. One limitation of previous TBLT research is that researchers did not investigate what the learners were thinking about when they engaged in the tasks. Asking the participants about their speaking performance immediately after finishing the task allowed me to better understand what they were thinking when making the monologic speech. The students first rated their perceptions of the difficulty of the monologue test using the following 4-point Likert scale: 1 = *Difficult*, 2 = *Relatively difficult*, 3 = *Relatively easy*, 4 = *Easy*. They also wrote why they chose the rating. Second, they answered what they had prioritized when completing the monologue test by slecting one of the following five options: focus on content, grammar, vocabulary, organization, and formulaic language. They chose one or multiple answers from the five choices. They then wrote descriptions of what they did in Japanese.

3/2/1 Training Reflection Questionnaire

The participants completed the Japanese version of the 3/2/1 Training Reflection Questionnaire (See Appendix H for the Japanese version and Appendix I for the English translation), in which they reflected on their perceptions of the 3/2/1 speaking tasks in Week 13. The purpose of this questionnaire was to gather information about the students' perceptions of the 3/2/1 task and the pedagogical interventions. The questionnaire was written in Japanese. The questionnaire was piloted with eight students attending the same university. The students in the pilot study understood the questions and completed the questionnaire appropriately.

The number of questions differed depending on the experimental group. First, questions about the 3/2/1 task were asked to all participants. The participants rated their performance on the 3/2/1 task by responding to the statement: *I am good at speaking in the 3/2/1 speaking tasks* using a 6-point rating scale: 1 = *Strongly disagree*, 2 = *Disagree*, 3 = *Slightly disagree*, 4 = *Slightly agree,* 5 = *Agree*, 6 = *Strongly agree*. The participants also provided reasons for the rating they chose by responding to an open-ended question. The participants answered three other open-ended questions asking what they prioritized when they talked for three minutes and one minute, and their comments on the 3/2/1 task.

The two experimental groups, the teacher-led planning group and the teacher and peer group, answered additional questions about their pedagogical intervention by responding to the following statements: *I think teacher model input is necessary, I use the teacher's model speech as a reference for 3/2/1 task*, and *I think peer-check activity is effective*. They responded by using the following 6-point rating scale: 1 = *Strongly disagree*, 2 = *Disagree*, 3 = *Slightly disagree*, 4 = *Slightly agree,* 5 = *Agree*, 6 = *Strongly agree*. In addition, they wrote why they chose their responses.

Interviews

I conducted follow-up interviews in order to triangulate and expand on the information obtained from the questionnaire. Four students volunteered and signed an agreement to participate in an interview: Sumi and Megu (Teacher and peer group), Kenta (Teacher-led group), and Nana (Comparison group). The interview questions are shown in Appendix J. All names are pseudonyms.

I conducted semi-structured individual interviews in a classroom on campus about their perceptions of their linguistic development and their experiences in the 3/2/1 task. The interview questions were based on the participants' answers to the 3/2/1 Training Reflection Questionnaire and the 3/2/1 Recording Retrospective Questionnaire. I asked the participants (a) whether and how their speaking had changed from the beginning of the semester to the end of the semester and why it had changed, (b) how useful the 3/2/1 task, the teacher-led model, and the peer check sheets were in helping them develop their language use and speaking ability such as, *Please tell your impression of the 3/2/1 task. Is it necessary to have thinking time (planning time) before doing the 3/2/1 task?*, and *What kind of things do you think about when planning?* The students answered in Japanese; their accounts were recorded and subsequently transcribed for analysis.

Mixed Method Research Design

In the social sciences, mixed method research has been increasingly recognized as a research paradigm that has arisen in response to the controversial issue of the dichotomy between qualitative and quantitative research methods (Hashemi & Babaii, 2013; Johnson, Onwuegbuzie, & Turner, 2007). Recently, mixed-method research has been also advocated by applied linguists (Brown, 2014; Hashemi & Babaii, 2013). Johnson et al. defined mixed methods research as a type of research in which a researcher or team of researchers combines elements of qualitative and quantitative research approaches (e.g., use of qualitative and quantitative viewpoints, data collection, analysis, inference techniques) for the broad purposes of breadth and depth of understanding and corroboration (p. 123). Johnson and Onwuegbuzie (2004) listed the strengths and weaknesses of mixed-method research (See Table 3).

One advantage of using mixed method in this study is that open-ended questionnaire responses and interviews allow me to better understand the participants' language development. Johnson and Onwuegbuzie (2004) stated that words, pictures and narratives can be used to add meaning to numbers. Examining qualitative data helps me to understand why a particular group or participants performed that way or what they were trying to do during the task performances. One disadvantage of conducting a mixed method study is that it is time-consuming to collect data and analyze interviews and open-ended questionnaires. Researchers must understand the right timing and appropriate way of gathering both qualitative and quantitative data; therefore, it can be challenging for researchers to complete all the data collection and analyses.

The rationale for using a mixed method design in this study was that the quantitative analysis of learners' oral performance provides a limited understanding of how learners perform on the 3/2/1 task and perceive the pedagogical interventions. For example, merely counting mean length of runs or clauses per AS unit cannot indicate what the learners prioritized during their performances and how they perceived the pedagogical interventions during the 3/2/1 task. Using qualitative data to illuminate the quantitative results allows for a more complete understanding of the quantitative results.

Table 3. *Advantages and Disadvantages of Mixed Methods (Johnson & Onwuegbuzie, 2004, p. 21)*

Advantages	Disadvantages
Words, pictures, and narrative can be used to add meaning to numbers.	It can be difficult for a single researcher to carry out both qualitative and quantitative research, especially if two or more approaches are expected to be used concurrently; it may require a research team.
Numbers can be used to add precision to words, pictures, and narratives.	Researchers must learn about multiple methods and approaches and understand how to mix them appropriately.
It can provide quantitative and qualitative research strengths.	Methodological purists contend that one should always work within either a qualitative or a quantitative paradigm.
Researchers can generate and test a grounded theory.	It is time consuming.
It can answer a broader and more complete range of research questions because the researcher is not confined to a single method or approach.	It is expensive.
Researchers can use the strengths of an additional method to overcome the weaknesses in another method.	Some details of mixed research remain to be worked out fully by research methodologists (e.g., problems of paradigm mixing, how to qualitatively analyze quantitative data, how to interpret conflicting results).
It can be used to increase the generalizability of the results.	

Three issues need to be considered when conducting a mixed method study: timing, weighing, and mixing (Cresswell & Plano Clark, 2007). First, I considered the timing of collecting data; a sequential design was used. The sequential design means that the quantitative data analysis occurred with the students' oral performances. In order to understand the participants' perceptions of the treatment, they had to finish the treatment before responding to the questionnaire items.

Second, weighing, which concerns the importance or priority of the quantitative or qualitative methods, is important because it indicates how the research is conducted and the theoretical perspective used when investigating the research questions. In this study, quantitative analyses played the major role and qualitative analyses, which were used to supplement and illuminate the quantitative results, played a secondary role.

Third, it is necessary to consider how to combine the quantitative and qualitative methodologies. In this study, I connect the data between two phases. After analyzing the quantitative data, I looked at the students' 3/2/1 retrospective questionnaire responses in order to investigate how the participants perceived the intervention and the recording. Cresswell and Plano Clark (2007) described this type of mixing as connecting data analysis to data collection by saying, "A researcher may obtain quantitative results that lead to the subsequent collection and analysis of qualitative data" (p. 84).

Cresswell and Plano Clark (2007) divided mixed methods designs into four categories: Triangulation Design, The Embedded Design, The Explanatory Design, and The Exploratory Design. The Explanatory Design, which uses qualitative data to explain significant results, outlier results, or surprising results (p. 72), was used in this study (See Figure 2). In this design, researchers collect quantitative data first and then collect qualitative data to explain or elaborate on the quantitative data. The rationale is that quantitative findings can provide a general answer to the research question, while qualitative analyses explain those statistical results by exploring participants' perceptions in greater depth. Therefore, the data can be described as QUAN + qual.

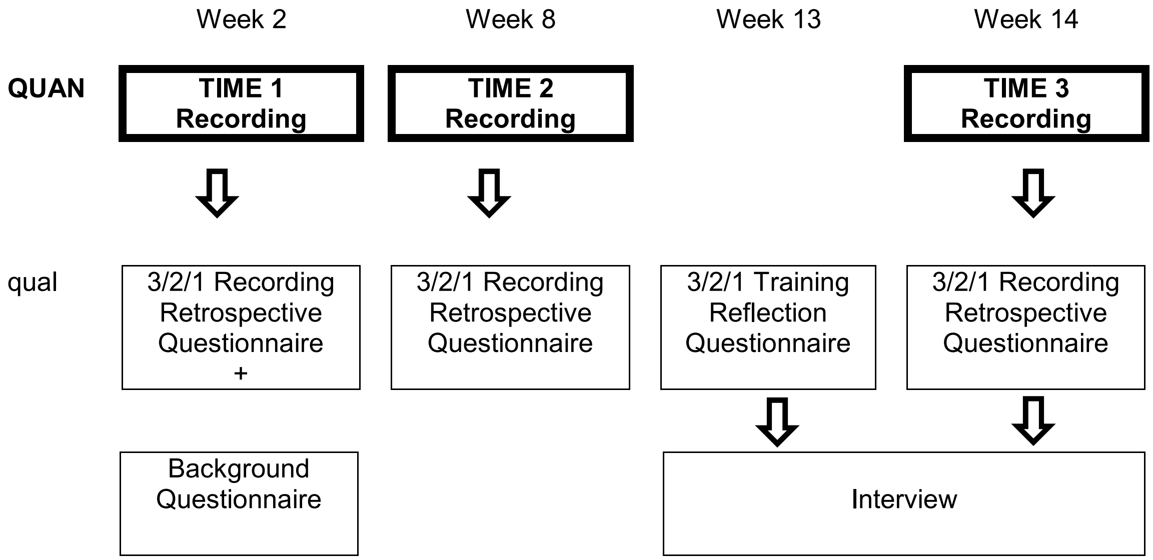

Week 2
Week 8
Week 13
Week 14
QUAN
TIME 1 Recording
TIME 2 Recording
TIME 3 Recording
qual
3/2/1 Recording Retrospective Questionnaire +
3/2/1 Recording Retrospective Questionnaire
3/2/1 Training Reflection Questionnaire
3/2/1 Recording Retrospective Questionnaire
Background Questionnaire
Interview

Figure 2. The explanatory research design.

Procedures

The 3/2/1 speaking task, was implemented in every class. Instead of the well-known 4/3/2 task, a shorter 3/2/1 task was used (e.g., Thai & Boers, 2016). In this task, one speaker talks about a particular topic for three minutes, retells the information a second time in two minutes, and then retells it a third time in one minute. This shorter 3/2/1 task was used for two reasons. First, the curriculum did not allow me to spend a longer time on this activity in each 90-minute class. Second, some participants might have been unable to continue speaking for four minutes due to their low oral proficiency. If the learners had been provided with too much time, it might have allowed them to rely on on-line planning rather than pre-task planning (Ogawa, 2016).

The 3/2/1 task was implemented in the beginning of every lesson because it was used as a warmer and lesson topic introduction. The 3/2/1 task questions used in this study were written in the in-house course textbooks. The 3/2/1 topics were relevant to the discussion topics that the students discussed in the day's lesson. Table 4 shows the target formulaic language training schedule and the 3/2/1 task questions for each week of the academic semester.

In the 3/2/1 task, the students formed pairs of speakers and listeners. Speakers spoke on the topic for the allotted time, and the listeners did not interrupt the speakers by making comments or asking questions. The speakers completed a full 3/2/1 task with three different listening partners each round before switching roles. For example, the speaker worked with Partner A for three minutes, Partner B for two minutes, and Partner C for one minute. Between rounds, the speakers were told to provide the same information in a reduced time. When there was an odd number of students, I joined the 3/2/1 task as a listener. When the first 3/2/1 set was completed, the students changed roles, and I exited the activity leaving one speaker with two listeners. In this way, I could observe the other students and not take a speakers' role in the task.

Explaining and Practicing the 3/2/1 Training (Week 1)

In the first week of the semester, the participants received a teacher-fronted explanation about how to engage in the 3/2/1 speaking task training. The purpose of this explanation was to ensure that the students understood the procedures of the 3/2/1 task correctly. I emphasized the following three points. First, the speakers should repeat their talk as the time becomes shorter on the second and third trials. As the time becomes shorter, they should not delete information nor should they add new information. Second, the speakers must speak faster in order to repeat everything within a shorter time span.

After I finished explaining the 3/2/1 task procedure orally, the students practiced the 3/2/1 task with a partner using the following three questions; *Do you like your hometown? Why did you choose your department or this university?* and *What do you do in your free time?* Because it was the students' first time to attempt the 3/2/1 task in the first English discussion class, I provided the students with three minutes of planning time. The students were instructed to brainstorm using only English words and not to write sentences on a blank piece of paper.

Because this was the students' first opportunity to practice the 3/2/1/ task, the main purpose was to familiarize them with the task procedures. Therefore, the students could look at the brainstorming handout while they talked. No target formulaic language was presented. Most of the students were able to speak for the full 3 minutes by looking at the handout. They also tried to produce the same information faster in the second and third iterations.

3/2/1 Treatment Phase (Weeks 3-13)

From the third week, the three groups of participants engaged in 3/2/1 training with different pedagogical interventions. Each group was made up of two or three intact classes. I randomly sorted the class combinations based on their TOEIC scores so that the participants' English proficiency was similar in each group.

Table 4. *The Target Formulaic Language Training Schedule and the 3/2/1 Task Questions*

Week	Lesson topic	Formulaic language taught	3/2/1 questions (Treatment phase and data collection)
1	Practice and introduction		• Do you like your hometown? Why? • What do you usually do when you have free time? • Why did you choose this department of this university?
2	Time 1 (Recording)	Opinion • *In my opinion* • *Personally speaking, I think* • *I am not sure but I think*	Time 1 Questions (Recorded) • *Do you think doing club activities is a good idea for students? Have you ever joined a club before? What did you learn from your experiences? Why did you choose your club in this university?*
3	Communicat ion	Reasons • *It's mainly because* • *One reason is* • *Another reason is*	• What did you enjoy doing with your friends in your high schools? • What do you enjoy doing with your friends at university?
4	Education		• Do you think going to university is important? Why did you decide to come to university? • What are your future plans after you graduate from university? (e.g., job, marriage, travel?)
5	Education		• Do you think cram schools are important? • Which is better, entrance exam system or recommendation system to get into university?
6	Environment	Examples • *For example* • *For instance* • *One example is* • *Another example is*	• Do you think Japanese people are eco-friendly? Are you eco-friendly? How about your family or part-time job? • Do you think Tokyo is clean?
7	Environment		• Do you think your hometown is a nice place to live? • Where would you like to live in the future?
8	Time 2 (Recording)		Time 2 Questions(Recorded) • *Do you think eating out is better than eating in? Do you often eat out or eat in? What kind of food do you like? Who do you usually eat dinner with?*

Table 4 (continues).

Table 4 (continued).

Week	Lesson topic	Formulaic language taught	3/2/1 questions (Treatment phase and data collection)
9	Social issues		• Do you think you are independent from your parents? Do you think being independent 10is important? • Do you have pressure from your parents?
10	Technology		• Do you think technologies are important for you? • Which technologies do you want to buy?
11	Technology	Possibility • *If*	• Do you think SNS is good for you? What websites, apps, and SNS do you often use?
12	Values		• What makes you happy? What made you happy when you were younger? • What makes you unhappy?
13	Values		• Which is more important, love or money? • What are important values for children to learn?
14	Time 3 (Recording)		Time 3 Questions (Recorded) • *Do you think learning English is important for students? Do you think study abroad is a good idea for students? What are other good ways to improve your English skills?*

The comparison group immediately started the 3/2/1 speaking task. They had no teacher-led planning nor were they instructed to use the target formulaic language during the task. The teacher showed the day’s topic (e.g., *What did you enjoy doing with your friends in your high schools? What do you enjoy doing with your friends at university?*), which was written on the paper. The participants were instructed to form pairs and decide which person would be the first speaker. The teacher told the first speaker to talk about the given topic for three minutes and told the listeners not to interrupt them by making long comments or by asking questions. The teacher set a timer for three minutes and put it in the front of the classroom so that everyone could see the time. After the teacher said “start,” the speakers began talking. It usually took 2-3 minutes between the teacher’s announcement of the topic and students' start of the 3/2/1 task. Ethically, this approach was acceptable because the control group engaged in a typical 4/3/2 task.

Before the participants engaged in the 3/2/1 task, I encouraged the two experimental groups to use reasons and examples to support their opinions by showing a model on the whiteboard (see Figure 3). This model allowed the students to understand why these reasons, examples, and expression of possibility would effectively support their opinions.

How to stretch your speaking

OPINION → *REASON 1* → EXAMPLE 1-a → EXAMPLE 1-b

→ *REASON 2* → EXAMPLE 2-a → EXAMPLE 2-b

Figure 3. The monologue model shown on the white board.

The two treatment groups received a pedagogical intervention that involved the students in using the target formulaic language during the 3/2/1 task. The target formulaic language, which is listed in Table 4, was the same functional phrases that were taught in the main part of the class (outside of the treatment phase) for the group discussion. The target formulaic language was introduced to all the students outside of the 3/2/1 training phrase during the lessons because the phrases were introduced in the course textbook as part of the group discussion task.

The teacher-led planning group received the teacher-led model passage using the formulaic language with the handout (Appendix K) prior to engaging in the 3/2/1 speaking tasks each lesson. The teacher-modeled passage was displayed on the handout with the target formulaic language, which was enhanced with underlining.

All types of the target formulaic language (e.g., opinion, reason, example, possibility) were provided in the model input, and then the teacher-led and the teacher and peer group read the teacher-led model; thus, although the students did not study some of the target formulaic language until later in the semester (e.g., the students were introduced to the example function in Week 6 and the possibility function in Week 11), they were exposed to the teacher-modeled passage earlier during the 3/2/1 task. While the teacher read aloud the passage, the students read the passage silently. Reading the teacher-modeled passage took approximately 1 minute (see Table 5). After the participants finished reading the teacher-modeled input, they were given two minutes to plan what to say. They did this by writing phrases in English on the handout to generate ideas. The participants were not allowed to look at the handout when they performed the 3/2/1 task. I also encouraged the participants to organize their speech by supporting their opinions with reasons and examples so that they would be able to speak longer. The total length of the treatment was 3 minutes (1 minute of reading teacher-modeled passage and 2 minute of pre-task planning).

The teacher and peer group received an additional pedagogic intervention. While the participants engaged in 3/2/1 task, pressure to use the target forms was provided by the listeners. This intervention was designed to encourage the speakers to use the target formulaic language. The listeners determined whether the speaker was using the target formulaic language using a pair check card (Appendix L). Accordingly, the teacher and peer group followed the following peer-check intervention schedule; opinion (Week 2), reasons (Week 3), example (Week 6), and possibility (Week 11) (Table 4, Appendix L). The participants in the teacher and peer group received 3/2/1 task training in which they practiced opinion phrases for 12 weeks, reason phrases for 11 weeks, example phrases for eight weeks, and a possibility phrase for three weeks as they engaged in peer-check training. The total length of the treatment was 15 minutes (1 minute of reading teacher-modeled passage, 2 minute of pre-task planning, 12 minutes for checking their speaker partner to use the target formulaic language) and being checked by their listener partner. Table 5 shows the pedagogic intervention for each group.

Table 5. *Pedagogic Intervention*

Time	Comparison group (*n* = 12)	Teacher-led group (*n* = 13)	Teacher and peer group (*n* = 21)
Display the topic (30 seconds)	Teacher shows today's 3/2/1 task topic.	Teacher shows today's 3/2/1 task topic.	Teacher shows today's 3/2/1 task topic.
Text enhancement (1 minute)	None	Students read the teacher-modeled passage.	Students read the teacher-modeled passage.
Pre-task planning (2 minutes)	None	Students write their ideas on the back side of the paper.	Students write their ideas on the back side of the paper.
(13-15 minutes)	Students do the 3/2/1 task.	Students do the 3/2/1 task.	students do the 3/2/1 task with peer check .

3/2/1 Recording Procedure (Weeks 2, 8, 14)

The students' oral performances during the 3/2/1 task were recorded three times in Weeks 2, 8, and 14. The students recorded their 3/2/1 task production individually on the test days. The 3/2/1 individual monologue was selected for the following reasons. First, testing the students' speaking ability using the 3/2/1 task was appropriate because the participants were familiar with the task from the in-class instruction; as a result, they were able to focus on the linguistic elements of the test and not on the test procedures.

Second, individual monologues made it possible to control the order effects. The participants performed the 3/2/1 task in pairs during the treatment stage. If the participants were allowed to use the same procedure during the recording, the second speaker would have had an advantage of listening to their partners and receiving their partners' input before speaking. If different questions had been provided to the first and the second speaker, the problem of using different topics among the participants would have arisen. Therefore, providing the participants the same questions was the best option.

In order to assess the development of speaking proficiency, similar types of questions were used in all three tests (Table 6). In order to control for topic repetition effects, different questions were used at each data collection stage. The test questions were created with the following points in mind: (a) the speakers needed enough questions to continue talking for three minutes, (b) the topic needed to be familiar but it should not have been used in the courses previously and (c) the topics needed to be appropriate for the participants' proficiency level. In order to ensure that the 3/2/1 recording questions were appropriate for the students' English proficiency levels, the same questions were piloted with eight students who were not in the study prior to test administration. The students in the pilot study were able to understand the questions clearly and they could answer these questions properly. Time on task was also appropriate because many students were able to continue speaking for the allotted time. To ensure that the participants understood the questions, a Japanese translation was provided under the

English questions on the recording days at Time 1 (Appendix M), Time 2 (Appendix N) and Time 3 (Appendix O).

Table 6. *3/2/1 Recording Questions*

Test	Questions
Time 1 (Week 2)	Do you think doing a club activity is a good idea for students? (Have you ever joined a club before? What did you learn from your experiences? Why did you choose your club in this university?)
Time 2 (Week 8)	Do you think eating out is better than eating in? (Do you often eat out or eat in? What kind of food do you like? Who do you eat dinner with?)
Time 3 (Week 14)	Do you think learning English is important for students? (Do you think studying abroad is a good idea for students? What are other good ways to improve your English skills?)

In Week 2, the individual 3/2/1 monologue speaking task was recorded. The recording was made in the middle or at the end of the class for two reasons. The first reason was to have a trouble-free administration of the recording. One disadvantage of making the monologue recording at the beginning of the class was that participants sometimes entered the classroom late; they might have distracted other students while they were recording. The second reason was to minimize the participants' anxiety. The participants might have been anxious or not warmed up enough to speak English in the beginning of class, particularly, given that most of the classes were held at 9:00 am. In addition, some participants might have felt anxious about attending English discussion classes conducted in English, especially when the first recording was made (Week 2). Providing an IC recorder might also have made the students anxious at this early stage of the semester. For these reasons, the recordings were made in the middle or at the end of the lesson.

Prior to recording, the participants were informed about the study and asked to sign the consent form (Appendix A). I also informed the participants that their recordings were not related to their course grades. The following points were emphasized in the task instructions. First, speakers should talk as much as possible for the full time. Second, the speakers should talk about the same information in the second and third iterations without deleting or changing the content. Third, in order to repeat everything that they said in the first iteration, the speakers should speak faster in the second and third iteration.

After the instruction, the 3/2/1 recording questions were shown to the students on a hand out, and a written Japanese translation was provided to ensure that everyone understood the questions (Appendix M). One-minute of planning time was then provided because it allowed the participants to prepare for the 3/2/1 recording and it helped them to continue speaking for the entire time. When the participants were planning, they were instructed to brainstorm ideas on a handout for one minute. For example, the participants thought about why they thought club activities are important. After one minute of planning, the handouts were collected so that the participants could not look them while speaking. Each participant then moved to a different desk, which was facing a wall in the classroom. Approximately 1.5-2 meters was provided between the desks so that each participant would not be disturbed by the other students. During the 3/2/1 recording, each participant sat on the chair and recorded their monologue. The 3/2/1 monologue recordings were made using the same procedures in Week 8 (Time 2 recording) and Week 14 (Time 3 recording).

Data Coding Procedure

Transcribing

Only the two-minute performance of the recorded 3/2/1 speaking data was analyzed for two reasons. First, two-minutes is considered as an appropriate length, as it includes enough data to capture L2 learners' oral development. Two-minute speeches are often used in speaking tests such as the monologue task on the STEP *Eiken* first level test, which is a high-stakes English proficiency test used widely in Japan. In addition, the participants might have been able to most effectively show their oral speaking development in the two-minute performance. Speaking for 3 minutes might have resulted in the participants pausing frequently because it was the first performance. Speaking for 1 minute might not have allowed the students to use many exemplars of the target formulaic language.

Second, the raters were able to assess the participants' speech organization more efficiently in the two-minute performances than the three-minute performances. If the raters had listened to the participants' three-minute performances, it might have caused fatigue because they would have had to listen to many speeches. In addition, it might have been difficult for the raters to simultaneously rate four components—organization, complexity, accuracy and fluency—in a one-minute speech.

The speech data were transcribed in the following manner. First, I transcribed the recorded data including fillers and self-repetitions. At this time, pause length was not included. Next, the transcriptions were double-checked by a research assistant, a Japanese woman who had studied abroad in the United States for one year. When she heard additional sounds or different words, she put them in brackets, I then checked the speech data once again to confirm the changes to the original transcript. A total of 288 minutes of speech data were transcribed (2 minutes x 48 participants x 3 times). The original transcription was used when making the pruned speech transcription (see the next section), marking AS-unit boundaries and clauses, and measuring pauses using the PRAAT software (see the CAF analysis section).

After the speech samples were transcribed, transcriptions of the pruned speech, in which fillers, self-corrections, and repetitions were excluded, were produced in order to analyze syntactic complexity and accuracy. The rational for using pruned speech when measuring syntactic complexity was to avoid measuring complex sentences incorrectly. Longer utterances were considered more complex. However, if repetitions are included when measuring syntactic complexity, an utterance such as, *I'm think… I was think ..I was thinking…* is as complex as the utterance *I think my cat is very old, too,* both of which consist of eight words. Pruned speech was also used for calculating syntactic accuracy to take account of self-corrections after the speakers noticed syntactic errors. For example, if a speaker made a self-correction such as "She {have}…. has," it was counted as a correct utterance because the speaker noticed the error and self-corrected; pruned speech avoids the possibility of decreasing syntactic accuracy measures.

AS-Units

AS-units (Analysis of Speaking units) were used to assess syntactic complexity and morphosyntactic accuracy (e.g., De Jong & Vercellotti, 2016; Foster et al., 2000; Lambert, Kormos & Minn, 2017; Li et al., 2015; Nitta & Nakatsuhara, 2014; Révész et al., 2016; Tavakoli, Campbell & McCormack, 2016). Foster et al. defined an AS-unit as “a single speaker's utterance consisting of an independent clause, or sub-clausal unit, together with any subordinate clauses associated with either” (p. 365). Foster et al. stated that AS-units are better at capturing aspects of spoken language that other units might miss or categorize as errors. According to Foster et al., the following are examples of AS-units. AS-unit boundaries are marked by an upright slash (|), a clause boundary with an AS-unit is marked by a double colon (::), repetitions, false-starts, and reformulations are placed inside of brackets { } to be a part of the same AS-unit. Pauses are placed inside of the brackets ( ). The following bullet points indicate the types of categorizing AS-units.

- An independent clause is minimally a clause including a finite verb. Finite clauses must contain a verb that shows tense.

  Example:

  | English is important |

  | I spoke to my family yesterday |

  | My sister did not like the food |

- An independent sub-clausal unit consists of: either one or more phrases or more phrases which can be elaborated to a full clause by means of recovery of ellipted elements from the context of the discourse. The following examples were frequently observed in a dialogic task.

  Example:

  | How’s the weather today |

  | Raining |

  or irregular or non-sentence.

| Thank you very much |

| Good for you |

| Yes |

- A subordinate clause consists minimally of a finite or non-finite verb element plus at least one other clause element (subject, object, complement, or adverbial). Non-finite verbs typically mean a verb that is not influenced by tense or subject (e.g., He wants <u>to play</u>).

  Example:

  | I want :: to ask my friend :: if she can help me |

  | and I {I} am surprised :: that store will be closed |

  | I worked in a restaurant :: which has many foreign staffs |

- Coordinated clauses are considered to belong to the same AS-unit unless the first phrase is marked by falling or rising intonation and is followed by a pause of at least 0.5 seconds. In this study, because the participants have low-intermediate speaking proficiency, pauses more than 1 second before or after *and* or *but* were considered a different AS-unit (e.g., Kanda, 2015).

  Example:

  | I feel :: I'm enjoying my university life (1.2) | but I miss my high school friends |

  | I am trying {some} {my} my best (1.0) | and I also enjoy my club activity |

Syntactic Complexity

According to Norris and Ortega (2009), syntactic complexity should be measured multidimensionally. Length of AS-unit and amount of subordination are the main ways to measure syntactic complexity. In this study, syntactic complexity was measured using (a) mean length of clauses (pruned speech) (numbers of words/AS-unit) and (b) clauses per AS-unit (pruned speech). When calculating both complexity measurements, pruned speech was used. For (a) clauses per AS-unit, the amount of subordination was calculated by counting all clauses and dividing them by the number of AS-units.

Clauses can be categorized into independent and dependent clauses. Independent clauses can stand alone as a sentence, while dependent clauses cannot do so. Dependent clauses usually contain subordinating conjunctions (e.g., because, when, after, before, as, if), relative pronouns (e.g., who, whom, that, which), or nominal conjunctions (e.g., whether, that). In addition, clauses can be categorized into finite and non-finite clauses. Although there were many types of clauses, the purpose of this study is not to categorize the clause types. Therefore, I counted a clause as long as it was an independent clause, or a dependent clause that included both finite and non-finite verbs. The following are examples of how clauses were counted in this study. A clause boundary is marked by a double colon (::). For example:

| In my opinion, I think :: eating in is {more good} better than eating out :: because it is cheap | (3 clauses / 1 AS-unit)

| If I become rich :: I want :: to eat out | (3 clauses / 1 AS-unit)

| I like food :: which my mother make | (2 clauses / 1 AS-unit)

| Yes | (1 clause/ 1 AS-unit)

| For example, *sushi* and *udon* | (1 clause/ 1 AS-unit)

Non-finite clauses such as *I want to*, *I decide to*, and *It is important to do* were also counted. For example:

| I wish :: to er go to a nice restaurant in Tokyo | (2 clauses / 1 AS-unit)

| It is important for me :: to speak English | (2 clauses / 1 AS-unit)

| He decided :: to go shopping with his mother | (2 clauses / 1 AS-unit)

(b) Mean length was calculated by dividing the total number of words by the number of AS-units. For example:

| In my opinion, I think :: eating in is {more good} better than eating out :: because it is cheap| |I think :: I often eat out (1.2)| |but I sometimes feel tired of eating by myself | (31 words / 3 AS-unit = 10.33 words/ AS-unit)

Lexical Complexity

The lexical analysis involved calculating lexical diversity and the frequency of the usage of the target formulaic language. Recently, the importance of lexical analysis has been emphasized, as it contributes to a better understanding of the psycholinguistics of second language speech production and the inter-relationships among the CALF components (Skehan, 2009, p. 512). The lexis-syntax connection is vital in performance areas and vocabulary has a strong association with fluency (Skehan, 2009, p. 514).

Lexical diversity. One focus of this study was on the participants' lexical development during the one-semester treatment. Lexical diversity is often used to assess lexical growth (e.g., Crossley, McNamara, & Salsbury, 2009) because it is an appropriate lexical measure when analyzing short texts. In this study, the measure of textual lexical diversity MTLD (McCarthy & Jarvis, 2010) (http://textinspector.com/workflow) was used as the lexical diversity measure because it is the least sensitive to text length when the text analyzed consists of at least 100 tokens (Koizumi, 2012).

MTLD is an index of a text's lexical diversity evaluated sequentially. It is calculated as the mean length of sequential word strings in a text that maintain a given Type Token Ratio value (McCathy & Jarvis, 2010, p. 384). During the calculation process, each word of the text is calculated sequentially for its type-token ratio. For example, with Lincoln's speech, *of* (1.00) *the* (1.00) *people* (1.00) *by* (1.00) *the* (.8000) *people* (.667) *for* (.714)… and so forth. However, when the default type-token ratio factor size value (.720) is reached, the factor count increases by a value of 1, and the type-token ratio evaluations are reset (McCathy & Jarvis, 2010, p. 384). Therefore, MTLD would be carried out as *of* (1.00) *the* (1.00) *people* (1.00) *by* (1.00) *the* (.8000) *people* (.667) ||| FACTORS = FACTORS + 1 ||| *for* (1.00) *the* (1.00) *people* (1.00) and so forth (McCathy & Jarvis, 2010, p. 384).

Frequency of the target formulaic language. The raw frequency of the use of the target formulaic language per monologue speech was analyzed (Appendix E). The frequency was calculated in order to assess to what extent the students transferred their use of the target phrases from their treatment. Self-corrections and repetitions were not counted. The following utterances are examples of how the formulaic language was counted in the one-minute monologue tests: *I think….In my opinion….school uniforms are a good ideas. One reason is …one reason is…. I think, it is easy to choose. For example, if I have a school uniform, I can choose what to wear quickly every morning.*

In this case, the speaker self-corrected, *I think…in my opinion*; this utterance was counted as one instance of expressing an opinion. The speaker then stated a reason by repeating *One reason is…one reason is…*. This utterance was counted as one occurrence of expressing a reason. The speaker then said, *For example, if…*, which was counted as one instance of stating an example and one instance of stating a possibility because the speaker combined the two functions in a single utterance.

Morphosyntactic and Lexical Accuracy

Global accuracy was assessed but with specific notions after pruning. Morphosyntactic and lexical accuracy refers to the ability to avoid morphosyntactic errors (Ellis, 2009b; Skehan & Foster, 1999). The rational for using the global accuracy measure is that it was more applicable for the treatments in this study, given that no specific grammatical forms were taught; therefore, measuring specific grammatical errors (e.g., past tense, articles, or relative clauses) would likely show in little or no change. In addition, the participants rarely made mistakes on the target formulaic language because they were retrieving pre-fabricated, memorized chunks. Therefore, calculating errors in the use of the target formulaic language was not suitable, either.

Errors can occur with inflectional morphemes (e.g., third person singular *-s*, plural *-s*), function words (e.g., articles, prepositions), content words (e.g., adjective-noun collocations) and Japanese use (e.g., *igirisu* for England). Moreover, sometimes an utterance did not make sense; therefore, the error type could not be determined. The following are examples of errors the participants made.

Errors with inflectional morphemes:

| make many friends is very important in school lives | (*making)

| all great people is very kind to me | (*are, subject-verb agreement)

| I'm joined a basketball club | (misuse of be verb)

Errors with function words:

| this experience encouraged to me | (misuse of preposition)

Errors with content words:

| I grew up myself in the high school club | (misuse of the verb-noun collocation)

| I learned *ojigi* | (*bowing, Use of Japanese content words)

The entire AS-unit does not make sense

| less time is staying home holiday | (The sentence does not make sense)

| if circle or club is a chance in life | (Fragment)

Errors with vocabulary

| I can speak *graduately* | (Misuse of adverb gradually)

When analyzing oral data, it was sometimes difficult to judge whether the learner had used the past tense inflectional morpheme or not (e.g., I talk to [tɔk tu]). In some cases, the past tense [t] was unreleased when a stop consonant was followed by a consonant (e.g., I talked to [tɔkt˚ tu]). In addition, it was sometimes difficult to hear the indefinite determiner (e.g., When I was a [ə] high school student…). Therefore, errors relating to articles or voiceless sounds were not counted as errors.

I calculated the error-free AS units as follows. First, I counted the errors in the transcription based on the criteria shown above. I then counted the total number of AS units and the number of AS units without errors for each recorded monologue. The error-free AS units for each speech was calculated as the number of error-free AS units / the total number of AS units.

Fluency

According to Tavakoli and Skehan (2005), measurements of utterance fluency can be categorized into three groups: breakdown fluency, repair fluency, and speed fluency. Breakdown fluency and repair fluency are related to disfluency, while speed fluency is an oral fluency measure because longer pauses (breakdown fluency) and more repetitions and repairs (repair fluency) make a speech sound more disfluent, while mean length of run and mean duration of syllable (speed fluency) make a speech sound more fluent. Five measures of fluency were used in this study: Mean length of pauses, number of repairs and repetitions, mean duration for syllable, mean length of run, and phonation time ratio. Mean length of pauses is related to breakdown fluency, number of repairs and repetitions to repair fluency, and mean duration of syllables to speed fluency. Mean length of fluent runs and phonation time ratio are a combination of breakdown fluency and speed fluency.

Mean length of pauses. There are two kinds of pauses: silent pauses and filled pauses. Silent pauses were defined as pauses longer than 300 ms (e.g., De Jong & Bosker, 2013; Thai & Boers, 2016) given that the participants' oral proficiency level was low-intermediate. Nonverbal fillers such as *uh*, *ah*, and *um* were also treated as pauses (e.g., De Jong & Perfetti, 2011; De Jong et al., 2013; De Jong et al., 2015); thus, *pauses* in this study includes both silent pauses and filled pauses.

Length of pauses was measured using the PRAAT speech analysis software (http://www.praat.org) (Boersma & Weenink, 2009), which functions by creating speech objects, labeling a waveform, analyzing intensity, sonogram, pitch, and duration. In oral fluency studies, PRAAT has often been used to measure pauses and speech rate (e.g., De Jong & Wempe, 2009).

I listened to the audio recordings and identified pauses using PRAAT. The beginning and end of each speech segment was located first using the PRAAT function "To textgrid (silences)" (De Jong & Perfetti, 2011, p. 545). It took approximately 30 minutes to detect the duration of pauses and calculate the speech rate in each 2-minute segment. I double checked the audio and identified voiced/unvoiced phonemes and fillers. It took approximately 80 hours to analyze breakdown fluency in the 144 speech samples. After the pauses were identified, the mean length of the pauses was calculated by dividing the total length of the pauses by the number of pauses.

Number of repairs. Repair fluency includes false starts, reformulations, and the repetition of words or phrases (Tavakoli & Skehan, 2005, p. 255). My intention was not to distinguish repetitions and reformulations. In this study, all false starts, reformulations, and repetitions of words or phrases were counted as the same categories of repairs. Fillers were not counted as repairs in the analysis of repair fluency in order to avoid an overlap with breakdown fluency, which includes fillers.

First, false starts are words left as incomplete clauses and followed by a new start involving different lexis and syntax (Witton-Davies, 2014). Two examples of false starts are shown below. Repair incidents are placed inside of brackets { }; they are part of the same AS-unit.

| {sh} she doesn't like cooking |

| {speci} specific category is |

Second, reformulations occur when speakers make a false start followed by an utterance that is similar to the initial utterance, except that the lexis or morphosyntax have been changed (Witton-Davies, 2014). Examples of reformulations are as follows:

| {the second no} the third |

| {she don't} she doesn't go out everyday |

| My father went to {German} Germany |

Third, repetition occurs when speakers repeat exactly the same words. Where repetition has a rhetorical effect (*very, very long*), or is considered to reflect normal usage (*No, no...*), it was not counted as a repair. Examples of the repetition of words or phrases are as follows:

| {I think} I think that |

| We need to {keep on} keep on fighting |

The frequency of repair incidents, which includes the three categories of false starts, reformulation, and the repetition of words or phrases, were counted in the two-minute recorded monologues. Repetition can be challenging to count because repeating *I am, I am* is not the same as uttering *I am, I am, I am, I am*. When a speaker repeated the same word or phrase such as *I*, it was counted as one repetition. For example, each of the following AS units contains one occurrence of repair.

| {I } {I} I think :: eating at home is healthy |

| {I think} {I think} I think :: eating at home is healthy |

| {it's hard} {it's hard} {it's too hard} it was too hard |

| Today {I am work} I am going to work at a restaurant |

| Today {I am work} {I am going to} I am going to work at a restaurant |

If the repeated phrases were separated from one another, they were counted as multiple occurrences. For example, the following utterance contains two occurrences of repair.

| Today {I am work} I am going to work {at a} at a restaurant |

Some researchers (e.g., Kormos, 2006) have stated that repairs should not be seen as an indicator of dysfluency without further investigation because they indicate monitoring processes in speech. For example, speakers make a self-correction when they detect that their output has been erroneous or inappropriate and this processing indicates a modification of the preverbal plan. Higher-level speakers use repairs to avoid communication problems such as rephrasing pragmatically appropriate utterances. Having said that, because lower-proficiency L2 learners' system of knowledge is typically incomplete and their production mechanisms are not fully automatic. For example, the following utterances are from a participant in this study. The speaker reformulates many times until he finally decides to say *when join club activity make many friends.*

(eh) {club activity} (eh) {because (eh) club active} (eh) {making} (eh) {make (eh) club activities} (eh) {make friends} (eh) {make friends} (eh) {make} (eh) when join club activity make many friends.

Therefore, in this study, considering the level of the participants, repairs are considered a type of disfluency.

Mean duration of syllable. According to De Jong et al. (2015), unconfounded measures need to be chosen to accurately measure speed fluency. As a measure separate from other disfluency components such as pauses and repairs, speed fluency was calculated as the mean duration of syllables, which was calculated as speaking time divided by the number of syllables produced. As mentioned in the literature review, this measure was used in previous studies because it is a measure of how fast a speaker speaks (e.g., Bosker et al., 2013; De Jong et al., 2013, 2015). When analyzing the mean duration of syllables, speaking time was used after excluding pauses.

Mean length of runs. Mean length of runs can be used as an indicator of procedualization if it is used with the mean length of pauses and phonation/time ratio (De Jong & Perfetti, 2011; Towell et al., 1996). A combination of speed and flow can be used to investigate the proceduralization and automatization of speech performances. De Jong and Perfetti proposed that the following three fluency measures are indicators of proceduralization: mean length of pauses (in seconds), the phonation/time ratio, and the mean length of runs. The phonation time ratio captures the proportion of the total length of utterances, including non-lexical filled pauses, to the total length of speech production. It reflects not only the number of pauses, but also the length of pauses. The mean length of runs is a measure of the average span of speech without pauses.

According to Towell et al., there are two reasons why these three measures are important when assessing proceduralization and automatization. First, speakers might pause the same number of times but vary the average length of the pauses, thereby reducing the total pausing time. This change would be evident in differences in the mean length of pauses. Second, speakers can pause for the same average amount of time on each occasion but vary in terms of the number of pauses. Fewer pauses normally give rise to an increase in phonation time ratio, as more time is spent speaking and less time is spent pausing (Towell et al., 1996, p. 93).

In sum, when proceduralization has occurred, speakers are able to produce longer runs; however, they might take more time engaging in on-line planning, which might result in longer pauses and longer fluent runs. If the mean length of runs increases, but the mean duration of silent pauses and phonation/time ratio are the same, the learners did not use online planning to produce longer runs. This result might indicate that encoding and sentence building have been proceduralized (De Jong & Perfetti, 2011, p. 539).

The mean length of run was calculated as the mean number of syllables produced in an utterance between pauses (total number of syllables divided by number of runs). A run is a fluent sequence between two silent pauses. Pauses were identified using a cut-off rate of 300 ms. The number of runs was calculated by adding 1 to the number of pauses. For example, if there were seven pauses, then there were eight runs: 7 +1 = 8. Then, the total number of syllables was divided by 8. Syllables were counted using the website http://www.syllablecount.com (Arczis Web Technologies, 2019).

Phonation time ratio. Phonation/time ratio was calculated as the total length of phonation time (time spent speaking) divided by the total response time a participant spent speaking (2 minutes). Specifically, the following procedures were used to calculate the measure. First, the total length and the number of pauses was determined using a cut-off rate of 300 ms. Phonation time was determined by subtracting the total time of silent pauses from the total response time (e.g., 120 seconds total – 30 seconds pause length = 90 seconds).

Table 7 shows the calculations for the CALF measurements. There are five fluency measures: mean length of pauses, number of repairs, mean duration of syllable, mean length of run, and phonation time ratio. Pauses include both silent and filled pauses.

Inter-Coder Reliability

Because the analysis of speech data can vary based on rater subjectivity, the data were analyzed by two raters. First, all the transcriptions were double-checked for accuracy by a research assistant. Second, I coded all the data. To ensure the reliability of the CALF measures, approximately 10% of the total sample size was calculated by a research assistant. Five participants' speech data from each recording (Times 1, 2 and 3) were randomly selected and then checked for reliability. Percentage agreements were calculated for the classification of student output into AS-units and clauses. Initially, the percentage agreement was 73.3% for AS units, 86.6% for clauses, and 80.0% for error-free AS-units. All coded transcripts were compared, discrepancies were discussed, and

Table 7. *CALF Measurements*

CALF	Type	Specific measure	Calculation
Complexity	Syntactic complexity	Clauses per AS-unit	Number of clauses/number of AS units
		Mean length of AS units	Number of words/number of AS units
Accuracy	Global measures	% of error-free AS-units	Number of error-free AS-units/total number of AS-units
Lexis	Lexical diversity	Measure of textual lexical diversity	
	Frequency of usage of the target form	Formulaic language	Raw number of formulaic language per speech
Fluency	Breakdown fluency	Mean length of pauses	Sum of pauses/number of pauses
	Repair fluency	Number of repairs	Total number of repairs
	Speed fluency	Mean duration of syllables	Spoken time/number of syllables
	Combination	Mean length of run	Total number of syllables/number of runs
		Phonation time ratio	Spoken time/total time

Note. Spoken time means phonation time spent speaking without silent pauses and fillers. Spoken time includes repairs. Frequency of usage of the target form was used to answer Research Question 3. Other nine measures of CALF were used to answer RQ 1, 2 and 5.

agreement was reached for every case. The data were rechecked and recorded so inter-rater agreement was 100%. Word count, syllable count, and lexis measured by MTLD were calculated using software on a website, so inter-rater reliability was not calculated. Table 8 shows the steps for measuring the acoustic CALF measures. Prunes speeches were used to measure error-free AS units (accuracy), mean length of AS units (complexity), and clauses per AS units (complexity).

Table 8. *Steps for Measuring the Acoustic CALF Measures*

Step	Procedure	Reliability
Transcription	Transcribed all the words	All transcriptions were double-checked by a research assistant.
Determining AS-unit	Divide the utterances by AS-units	10% of the data was checked by another researcher.
Pruning	Prune the speech by excluding self-repetitions, self-correction and false-starts.	
Accuracy	(With pruned speech) Analyze the percentage of error-free AS units with pruned speech.	10% of the data was double-checked by another researcher.
Mean length of AS units (Complexity)	(With pruned speech) Analyze the mean length of AS units by counting the number of words per AS unit of pruned speech.	Microsoft word used to count words.
Subordinate clauses (Complexity)	(With pruned speech) Count the number of clauses and divide by the number of AS units.	
Speed fluency	Count all syllables including fillers and repetition.	The syllable counter was used.
Repair fluency	Count the number of repairs and filled pauses	All repairs and fillers were checked by a research assistant.
Breakdown fluency	Count the number and duration of pauses using PRAAT.	

Human Ratings

In addition to the analytical CALF measures, I employed human ratings because using the analytical CALF measures alone does not evaluate the degree to which participants achieve the task goal. According to Pallotti (2009), communicative adequacy is the degree to which a student is successful in achieving the task goal (p. 596). In line with Pallotti's definition, I developed the following rating rubric (Table 9).

The goal of the 3/2/1 task was similar to Sato's (2011) idea that speakers can successfully tell a coherent and organized monologue with sufficient information. I did not completely separate communicative adequacy from linguistic competence because learners need to show linguistic competence to convey their meanings and achieve the communicative goal. In addition, it is useful to get confirmation for objective results from human raters. Révész et al. (2016) assessed communicative adequacy separated from linguistic competence, namely, functional adequacy, while communicative adequacy in this study include both organization and linguistic competence such as complexity, accuracy, and fluency~~.~~

First, the rating scale for organization was developed. A coherent, well-organized speech allows listeners to understand the message clearly. I developed the rubrics keeping in mind that the descriptor of each point on the rating scale should match the description of the other rating components.

Second, subjective ratings for the CAF variables were developed. Human ratings of CAF allow researchers to understand the CALF constructs from a subjective perspective (Segalowitz, 2010). Instead of using discourse analytic measures of CAF, human ratings of CAF shed light on communicative adequacy in new ways because when people communicate in daily life, they do not analyze speech by counting the frequency of fillers or measuring pause length; instead, they assess the speech through intuitive impressions of the speaker's linguistic ability.

Table 9. *Human Rating Rubric*

	Organization	Complexity	Accuracy	Fluency
5 Very successful	Speech is extremely well organized and very coherent with detailed information.	A wide range of variety of grammar is used. Attempts the use of coordination and subordination to convey ideas very often.	Grammatical errors are absent or very rare.	Speech is extremely smooth. Hesitations rarely occur, and they are very short.
4 Successful	Speech is fairly well organized and coherent.	Fairly wide range of grammar is used. Often attempts to use coordination and subordination to convey ideas.	Grammatical errors are rare.	Speech is fairly smooth. Hesitations very occasionally occur, and they are short.
3 Moderately successful	Speech somewhat well organized and mostly coherent.	A somewhat wide range of grammar is used. Occasionally attempts to use coordination and subordination to convey ideas.	Grammatical errors sometimes occur.	Speech is somewhat smooth. Hesitations occur occasionally, but they are sometimes lengthy.
2 Poor	Speech is not well organized and is somewhat incoherent.	A limited range of grammar is used. Mostly relies on single clauses and simple phrases.	Grammatical errors often occur.	Speech is disfluent. Hesitations are frequent and often lengthy.
1 Unsuccessful	Speech is very poorly organized and is incoherent.	An extremely limited range of grammar is used. Completely relies on single clauses and simple phrases.	Grammatical errors very often occur.	Speech is very disfluent. Hesitations are very frequent and sometimes very lengthy.

The rubrics for assessing complexity, accuracy, and fluency were modified based on Iwashita et al. (2001). I employed a five-point scale in order to decrease the raters' cognitive load (Nemoto & Beglar, 2014). In this study, the analytical ratings were divided into five levels: 1 = *Unsuccessful performance*, 2 = *Poor performance*, 3 = *Moderately successful performance*, 4 = *Successful performance*, 5 = *Very successful performance*. The descriptions of the five levels are shown in Table 9.

After the rubrics were developed, 10 raters were recruited. I acted as the eleventh rater. The raters held a Master's degree in Applied Linguistics and related fields and they were all university professors. Six raters were Japanese, one was Canadian, one was British, one was Australian, and one was Chinese. Rater training was provided for approximately 40 minutes. The purpose of the training was to allow the raters to understand the criteria used to assess each of the components—organization, complexity, accuracy, fluency—and the general evaluation standard. First, I explained the rating tasks and the rubric. The raters then listened to four sample performances and assessed the participants' oral performances using a handout (Appendix P). Each sample audio file was from a different experimental group and different test time. Next, the raters and I discussed their ratings and the reasons for them. After the training, they rated 20-40 speeches at their own pace at home.

Analysis

Research question 1 concerned the extent to which participants who receive form-focus pedagogical intervention (teacher-led planning and peer-check to use the target formulaic languages) develop their oral performance over 13 weeks. This research question was answered by conducting three sets of repeated-measures ANOVA for the comparison, teacher-led group, and teacher and peer groups. The dependent variables were the nine CALF measures (Table 7). The first set concerned differences at Time 1, Time 2, and Time 3 for the comparison group, the second set was focused on changes at Time 1, Time 2, and Time 3 for the teacher-led group, and the third set concerned differences at Time 1, Time 2, and Time 3 for the teacher and peer group.

Research question 2 concerned possible differences in the oral performances of the three groups. This research question was analyzed by conducting a one-way ANOVA to compare mean differences between the groups for Time 1 and ANCOVAs for Time 2 and Time 3. An ANOVA was used at Time 1 in order to investigate whether there were any significant starting differences among the three groups. The independent variable was group (3 levels: comparison, teacher-led, and teacher and peer) and the dependent variables were the nine CALF measures (Table 7). The covariate was the CALF measures at Time 1.

Research question 3 concerned the extent to which the participants who received form-focus pedagogic intervention used the target formulaic language in terms of frequency and variety across the 13 weeks. This research question was answered by counting the frequency and the variety of the target formulaic language types. These data were descriptively analyzed. I counted the frequency of the target formulaic language (e.g., *in my opinion*, *it is mainly because…* or *for example…*). The average usage of the target formulaic language per person was calculated by dividing the total number of occurrences by the number of participants because the number of participants in each group differed. In addition to the frequency, transcription data were also analyzed qualitatively to determine how the participants used the target formulaic language.

Research question 4 concerned the extent to which the participants who received the form-focus pedagogic intervention developed their communicative adequacy over the 13 weeks. The scores were statistically analyzed using multifaceted Rasch analysis with the FACETS program. A repeated-measures ANOVA was then run for each group (comparison, teacher-led, and teacher and peer group). The dependent variable was human ratings of the participants' oral performances measured in Rasch logits at Time 1, 2, and 3. Details about the Rasch analysis are presented in the next section.

Research question 5 concerned the strength of the relationship among the objective CALF measures and the human ratings for communicative adequacy. This research question was answered by calculating Pearson correlations. The FACETS measures were the communicative adequacy measure; correlation coefficients were computed between the nine analytical CALF measures and the FACETS measure.

Research question 6 asked what the learners prioritized during the 3/2/1 monologue recording. This research question was answered by administering a retrospective questionnaire and conducting interviews. The participants in each group indicated what they prioritized during the monologue recording immediately after completing each test. They indicated the reasons for their responses and what they had prioritized during the recording the 3/2/1 task (i.e., content, vocabulary, formulaic languages, grammar, or organization). Their questionnaire responses were analyzed descriptively. In addition, the participants' answers to the open-ended questions were categorized to better understand what they had prioritized during the 3/2/1 task recording.

Rasch Analysis

A Rasch analysis was conducted using FACETS 3.71.4 (Linacre, 2013) to evaluate the communicative adequacy of the participants' oral performances. The many-faceted Rasch model (Linacre, 2002) has often been used by researchers to assess oral performances assessed by raters. FACETS calibrates the examinees, raters, tasks, and the rating scales onto the same equal-interval scale called the logit scale, creating a single frame of reference for interpreting the results of the analysis (Eckes, 2005). The many-faceted Rasch model makes it possible to include additional performance test variables as *facets* and to assess the participants' performances based on a number of facets in the performance setting. For example, in this study the facets were person ability, task difficulty, rater severity, and rating category difficulty. Person ability was estimated while taking the effects of the other facets into account. In addition, the many-faceted Rasch model shows how well each level of each rating category functions by providing information about rating scale step difficulty.

One advantage of using Rasch measurement is that it is possible to eliminate severity differences across raters. An assumption is that raters can discriminate better performances from poorer performances, and not rate in a random way. The many-faceted Rasch model treats raters as experts who can evaluate performances based on their understanding of the rating scale and test-taker performances (Bond & Fox, 2015).

Another advantage is that researchers can determine whether the judges rate consistently or not. The Rasch model produces interval measures and provides fit statics for raters, participants, and the rating scale. Fit is defined as "the degree of match between the pattern of observed responses and the modeled expectations. This can express either the pattern of responses observed for a candidate on each item (person fit) or the pattern for each item on all persons (item fit)" (Bond & Fox, 2015, p. 310).

A third advantage is that multi-faceted Rasch measurement matches one purpose of this study, which is to assess communicative adequacy. In this study, the raters evaluated the participants' oral performances based on topic development, fluency, accuracy, and complexity using a five-point rating scale. When combined in the FACETS analysis, the four categories were considered as a measure of communicative adequacy.

CHAPTER 4

RESULTS

In this chapter, the results of the six research questions are presented. Research question 1 concerns the degree to which each group developed over 13 weeks based on the CALF measures. Research question 2 concerns group differences on the CALF measures. Research question 3 concerns the development of formulaic language, and research question 4 concerns the development of communicative adequacy, as assessed by human raters. Research question 5 concerns the relationship among the CALF measures and communicative adequacy, and research question 6 concerns what the participants attended to (i.e., form or meaning) during the 3/2/1 recording.

Research Question 1: The Development of Speaking Proficiency Across One Academic Semester

Research question 1 asked about the extent to which the comparison group, the teacher-led group, and the teacher and peer treatment group improved syntactic complexity, morphosyntactic accuracy, lexis, and oral fluency across 13 weeks. This research question was answered by conducting three sets of repeated-measures ANOVAs. A one-way repeated-measures was conducted with the factor being time (3 levels: Time 1, Time 2, and Time 3) and the dependent variables were the nine CALF measures (Table 7). The first set concerned changes at Time 1, Time 2, and Time 3 for the comparison group, the second set was focused on changes at Time 1, Time 2, and

Time 3 for the teacher-led group, and the third set concerned changes at Time 1, Time 2, and Time 3 for the teacher and peer group.

Repeated-measures ANOVAs were run separately for the three groups. A mixed design ANOVA was not used because there was no need to look at interaction effects in this study. Because there are nine dependent variables, a Bonferroni adjustment was made. The dependent variables were checked to ensure that they were not too strongly correlated. This was done by producing Pearson correlation coefficients. This analysis was conducted to ensure that none of the variables had a correlation coefficient > .90 (Tabachnick & Fidell, 2007, p. 90), a level that indicates multicolinearity. Multicollinearity is a concern because when there is high correlation between two or more dependent variables, it is statistically and logically redundant to include both variables in the analysis.

Table 10 shows the correlation coefficients for the variables at Time 1 for the three groups. No variables displayed multicolinearity. Clauses per AS unit and mean

Table 10. *Correlations Among the Dependent Variables at Time 1*

	1	2	3	4	5	6	7	8	9
1. Clauses/AS	—								
2. ML of AS	.70**	—							
3. Error Free AS	-.18	-.10	—						
4. MTLD	-.12	.13	.01	—					
5. ML of pauses	.06	.17	-.01	.05	—				
6. Repair	-.18	-.10	.00	-.11	.04	—			
7. MDS	.03	.06	-.05	-.11	.30*	-.21	—		
8. MLR	-.23	.05	.10	-.01	-.10	.12	-.30*	—	
9. PTR	-.07	.01	-.01	-.22	.13	.38**	.07	.60**	—

Note. ** Correlation is significant at < .01 (2-tailed). * Correlation is significant at < .05 (2-tailed). Clauses/AS = Clauses per AS unit; ML of AS = Mean length of AS units; Error Free AS = Error Free AS unit; MTLD = the measure of textual lexical diversity; ML of pauses = Mean length of pauses; Repair = Number of repair occurrences; MDS = Mean Duration of Syllable; MLR = Mean Length of Run; PTR = Phonation Time Ratio.

length of AS units, two measures of syntactic complexity, were highly correlated, $r = .70$, $p < .01$, while phonation time ratio and mean length of run, two measures of fluency, were moderately correlated, $r = .60$, $p < .01$.

Table 11 shows the Pearson correlation coefficients among the dependent variables at Time 2. The multicolinearity assumption was checked and met, as the variables displayed correlations between -.02 and .73 ($M = .14$) for Time 2. Clauses per AS unit and mean length of AS units were highly correlated at $r = .73$, indicating that the longer an AS unit became, the more clauses the AS unit included. Phonation time ratio and mean length of run were relatively highly correlated at $r = .67$; thus, speakers who produced more syllables in one fluent run increased the phonation time ratio. Phonation time ratio and repair were moderately correlated at $r = .51$; the more repairs the speakers produced, the more time they spent speaking. Given that repairs are counted as speaking time, the phonation time ratio might not be a good indicator of speaking fluency.

Table 11. *Correlations Among the Dependent Variables at Time 2*

	1	2	3	4	5	6	7	8	9
1. Clauses/AS	—								
2. ML of AS	.73**	—							
3. Error Free AS	- .39**	- .48**	—						
4. MTLD (Lexical diversity)	-.05	-.04	.11	—					
5. ML of pauses	.17	.18	-.23	.03	—				
6. Repair	-.01	.06	.02	-.04	-.17	—			
7. MDS	-.16	-.18	.05	-.13	.26	-.25	—		
8. MLR	.13	.14	.19	.06	-.35*	.31*	-.38**	—	
9. PTR	.24	.30*	.03	.03	.07	.51**	-.10	.67**	—

Note. ** Correlation is significant at < .01 (2-tailed). * Correlation is significant at < .05 (2-tailed). Clauses/AS = Clauses per AS unit; ML of AS = Mean length of AS units; Error Free AS = Error Free AS unit; MTLD = Measure of textual lexical diversity; ML of pauses = Mean length of pauses; Repair = Number of repair occurrences; MDS = Mean Duration of Syllable; MLR = Mean Length of Run; PTR = Phonation Time Ratio.

Table 12 shows the Pearson correlation coefficients among the dependent variables at Time 3. The assumption that no variables displayed multicolinearity was checked and met, as the correlation coefficients were between -.01 and .80 (*M* = .11) at Time 3. Clauses per AS unit and mean length of AS units were highly correlated at *r* = .80, indicating that the longer an AS unit became, the more clauses the AS unit included. In addition, phonation time ratio and mean length of run were relatively highly correlated at *r* = .61, indicating that when speakers produced more syllables in one fluent run, they also increased speaking time. Finally, phonation time ratio and repair were moderately correlated at *r* = .44; thus, the more repairs speakers produced, the longer they spoke. Lastly, mean length of run and repair were moderately correlated at *r* = .43; thus the more repairs speakers produced, the more syllables they produced in a fluent run.

In sum, some variables had moderately high correlation coefficients ranging between .69-.80, but none exceeded the .90 criterion indicating multicolinearity (Tabachnick & Fidell, 2007).

Table 12. *Correlations Among the Dependent Variables at Time 3*

	1	2	3	4	5	6	7	8	9
1. Clauses/AS	—								
2. ML of AS	.80**	—							
3. Error Free AS	-.16	-.20	—						
4. MTLD (Lexical diversity)	-.02	.12	.37**	—					
5. ML of pauses	-.03	.04	.07	-.01	—				
6. Repair	.22	.15	.00	-.18	-.15	—			
7. MDS	.13	-.01	.14	.07	-.15	-.15	—		
8. MLR	.03	.06	-.02	-.12	-.25	.43**	-.31*	—	
9. PTR	.33*	.40**	-.07	-.10	.03	.44**	-.05	.61**	—

Note. ** Correlation is significant at < .01 (2-tailed). * Correlation is significant at < .05 (2-tailed). Clauses/AS = Clauses per AS unit; ML of AS = Mean length of AS units; Error Free AS = Error Free AS unit; MTLD = the measure of textual lexical diversity; ML of pauses = Mean length of pauses; Repair = Number of repair occurrences; MDS = Mean Duration of Syllable; MLR = Mean Length of Run; PTR = Phonation Time Ratio.

A series of repeated-measures ANOVAs was conducted for the nine dependent variables. To avoid committing a Type I error, a Bonferroni adjustment was used by dividing the alpha level of .05 by the number of dependent variables used in each type of CALF measure. This adjustment provides a balance between the possibility of committing and Type I and Type II error. There were two measures for complexity, one for accuracy, one for lexis, and five for fluency; thus, the alpha level was set at .025 for complexity (.05/2), .05 for accuracy (.05/1), .05 for lexical diversity (.05/1), and .01 for fluency (.05/5). In the following section, the results are shown for each of the nine dependent variables.

Clauses per AS unit (Syntactic Complexity). Table 13 shows the descriptive statistics for the clauses per AS unit measures for each group at Time 1, Time 2, and Time 3. All groups increased clauses per AS unit from Time 1 to Time 3; however, the mean scores for Groups 2 and 3 decreased from Time 1 to Time 2 before ultimately increasing at Time 3.

Before conducting the repeated-measures ANOVAs, the assumptions of the analyses were checked for each group at each of the three times. First, univariate outliers were checked by converting the raw scores for the speech data to *z*-scores and checking for values > ± 3.29 (Field, 2013). No univariate outliers were found.

Second, normality was checked as shown by *z*-skewness and *z*-kurtosis. Normality were met with *z*-skewness and *z*-kurtosis statistics < |2.58|, however, one variable (Clause per AS unit for Time 2, the teacher and peer group) had *z*-skewness and *z*-kurtosis statistics > |2.58| (see Table 13). The assumption of normality was violated for the teacher and peer group at Time 2 (*z*-skewness = 3.86, *z*-kurtosis = 4.75) as shown in Table 13. Fifteen out of 21 participants produced less than 1.5 clauses per AS unit and only one participant produced more than two clauses per AS unit, which made the distribution positively skewed. Therefore, a log transformation was used, as Field (2013) stated that a log transformation is useful for adjusting positive skew, positive kurtosis, unequal variances, and a lack of linearity (p. 203). After conducting the log transformation, the *z*-skewness was 2.85 and *z*-kurtosis became 2.33; these figures met the < |3.29| criterion. Table 14 shows the descriptive statistics of transformed data for the teacher and peer group.

Table 13. *Descriptive Statistics for Clauses per AS Unit at Time 1, 2, and 3*

	Comparison group			Teacher-led group			Teacher and peer group		
	Time 1	Time 2	Time 3	Time 1	Time 2	Time 3	Time 1	Time 2	Time 3
Minimum	1.17	1.47	1.60	1.25	1.07	1.36	1.14	1.22	1.31
Maximum	2.00	2.55	3.00	2.00	2.00	2.67	2.17	2.46	2.50
M	1.62	1.78	2.13	1.53	1.48	1.80	1.75	1.52	1.81
95% CI	[1.42, 1.81]	[1.59, 1.97]	[1.90, 2.36]	[1.40, 1.65]	[1.32, 1.63]	[1.61, 2.00]	[1.62, 1.88]	[1.39, 1.65]	[1.67, 1.96]
SD	.32	.32	.37	.21	.27	.33	.29	.29	.32
Skewness	-.01	1.25	.88	1.09	.40	1.35	-.66	1.93	.40
SES	.62	.62	.62	.60	.60	.60	.50	.50	.50
z-skewness	-.02	2.02	1.42	1.82	.67	2.25	-1.32	3.86	.80
Kurtosis	-1.53	1.44	1.49	1.35	-.69	2.65	-.69	4.62	-.57
SEK	1.19	1.19	1.19	1.15	1.15	1.15	.97	.97	.97
z-kurtosis	-1.29	1.21	1.25	1.17	-.60	2.29	-.71	4.75	-.58

Note. SES = Standard error of skewness, SEK = Standard error of kurtosis, CI = Confidence Interval.

Table 14. *Transformed Data for the Teacher and Peer Group for Clauses per AS Unit for Complexity*

Teacher and peer group	Time 1	Time 2	Time 3
Minimum	.34	.09	.12
Maximum	.06	.39	.40
M	.24	.17	.25
95% CI	[.20, .27]	[.14, .21]	[.22, .29]
SD	.08	.07	.08
Skewness	-.90	1.43	.09
SES	.50	.50	.50
z-skewness	-1.80	2.85	.18
Kurtosis	-.20	2.26	-.97
SEK	.97	.97	.97
z-kurtosis	-.21	2.33	-1.00

Note. CI = Confidence interval.

Three repeated-measures ANOVAs—one for each group—were run to examine if there was a significant growth in clauses per AS unit at Times 1, 2, and 3. The first ANOVA was run with data from the comparison group. Mauchly's test indicated that the assumption of sphericity was met, chi-square = .46, p = .79. There was a significant time effect, $F(2, 24) = 11.10$, $p < .001$, partial $\eta^2 = .48$, so three paired-samples t-tests were run to conduct post hoc comparisons. The results are displayed in Table 15. There was a significant gain between Time 1 and Time 3 and between Time 2 and Time 3.

Table 15. *Paired Comparison Results for Clauses per AS Unit for the Comparison Group*

	t	*df*	*p*	Cohen's *d*
Time 1-Time 2	-1.57	12	.140	0.43
Time 1-Time 3	-4.60	12	.001	1.27
Time 2-Time 3	-2.93	12	.013	0.82

The next repeated-measures ANOVA was run with the teacher-led group. Mauchly's test indicated that the assumption of sphericity was met, chi-square = 2.39, p = .30. A significant time effect was also found for the teacher-led group, $F(2, 26) = 6.13$, p

= .007, partial η^2 = .32. Three paired-samples *t*-tests were used to make post hoc comparisons between the three times. The results are displayed in Table 16. There was a significant gain between Time 1 and Time 3 and between Time 2 and Time 3.

Table 16. *Paired Comparison Results for Clauses per AS Unit for the Teacher-Led Group*

	t	*df*	*p*	Cohen's *d*
Time 1-Time 2	.60	13	.56	0.28
Time 1-Time 3	3.00	13	.01	0.87
Time 2-Time 3	-2.73	13	.02	0.74

The third repeated-measures ANOVA was run with the teacher and peer group. Mauchly's test indicated that the assumption of sphericity was met, chi-square = 1.01, *p* = .62. There was a significant time effect, *F*(2, 40) = 5.52, *p* = .008, partial η^2 = .22. Three paired samples *t*-tests were used to make post hoc comparisons between the three times. The results are displayed in Table 17. There was a significant decline between Time 1 and Time 2, and a significant gain between Time 2 and Time 3.

Table 17. *Paired Comparison Results for Clauses per AS Unit for the Teacher and Peer Group*

	t	*df*	*p*	Cohen's *d*
Time 1-Time 2	2.86	20	.01	0.63
Time 1-Time 3	-.69	20	.49	0.14
Time 2-Time 3	-2.86	20	.01	0.61

The last repeated-measures ANOVA was run for the teacher and peer group using the transformed data. Mauchly's test indicated that the assumption of sphericity was met, chi-square = 1.01, *p* = .60. There was a significant time effect, *F*(2, 40) = 5.95, *p* = .005, partial η^2 = .23. Follow-up paired-samples *t*-tests were run. The results are displayed in

Table 18. Similar to the untransformed data, there was a significant decline between Time 1 and Time 2 and there was a significant gain between Time 2 and Time 3.

Table 18. *Paired Comparison Results for Clauses per AS Unit for the Teacher and Peer Group with Transformed Data*

	t	*df*	*p*	Cohen's *d*
Time 1-Time 2	2.94	20	.008	0.80
Time 1-Time 3	-.65	20	.522	0.26
Time 2-Time 3	-3.04	20	.006	1.07

Figure 4 shows the changes in clauses per AS unit for the comparison, teacher-led, and teacher and peer groups. The transformed data is displayed for the teacher and peer group. The comparison and teacher-led groups significantly increased clauses per AS unit from Time 1 to Time 3, and from Time 2 to Time 3, while the teacher and peer group did not significantly increase clauses per AS unit during the academic semester.

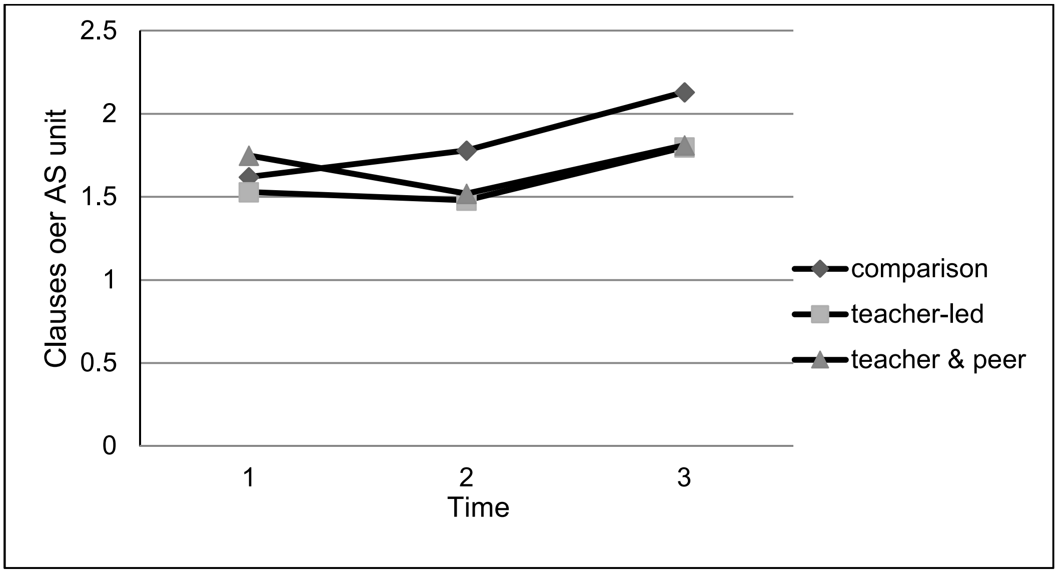


Figure 4. Changes in the clauses per AS unit at Times 1, 2, and 3.

Mean length of AS units (complexity). Table 19 shows the descriptive statistics for mean length of AS units for each group for Time 1, Time 2, and Time 3. The comparison group increased from Time 1 to Time 3, while the teacher-led group and the teacher and peer group decreased at Time 2 and improved at Time 3. All three groups had the highest mean scores at Time 3.

Before conducting the repeated-measures ANOVAs, the assumptions of the analyses were checked for each group at each of the three times. First, univariate outliers were checked by converting the raw scores for the speech data to *z*-scores and checking for values > ± 3.29 (Field, 2013). No univariate outliers were found. Second, normality was checked using the *z*-skewness and *z*-kurtosis criterion of < |2.58| (Table 19).

Three repeated-measures ANOVAs were run to investigate whether there was significant growth in mean length of run for the three groups. The first repeated-measures ANOVA was run with data from the comparison group. Mauchly's test indicated that the assumption of sphericity was met, chi-square = 2.29, $p = .31$. There was a significant difference, $F(2, 24) = 12.17$, $p < .001$, partial $\eta^2 = .50$, so three paired samples *t*-tests were run to make post hoc comparisons among the three times. The results are displayed in Table 20. There was a significant gain between Time 1 and Time 3 and between Time 2 and Time 3.

Table 19. *Descriptive Statistics for Mean Length of AS Units for Times 1, 2, and 3*

	Comparison group			Teacher-led group			Teacher and peer group		
	Time 1	Time 2	Time 3	Time 1	Time 2	Time 3	Time 1	Time 2	Time 3
Minimum	7.17	7.92	11.31	7.71	6.67	7.89	7.70	7.22	7.88
Maximum	13.86	14.55	17.00	11.67	15.00	15.22	14.00	11.54	14.14
M	10.62	10.81	13.63	10.02	9.57	11.43	10.86	9.38	11.15
95% CI	[9.40, 11.83]	[9.60, 12.00]	[12.50, 14.76]	[9.30, 10.74]	[8.30, 10.84]	[10.27, 12.59]	[10.03, 11.68]	[8.80, 9.96]	[10.29, 12.01]
SD	2.01	1.99	1.87	1.25	2.20	2.00	1.80	1.27	1.88
Skewness	0.15	0.28	0.56	-0.34	1.18	0.43	0.04	-0.09	-0.03
SES	0.62	0.62	0.62	0.60	0.60	0.60	0.50	0.50	0.50
z-skewness	0.24	0.45	0.91	-0.57	1.98	0.72	0.08	-0.18	-0.06
Kurtosis	-0.80	-0.23	-0.92	-1.20	1.58	0.12	-0.87	-1.00	-0.88
SEK	1.19	1.19	1.19	1.15	1.15	1.15	0.97	0.97	0.97
z-kurtosis	-0.67	-0.20	-0.77	-1.04	1.37	0.10	-0.90	-1.03	-0.90

Note. SES = Standard error of skewness, SEK = Standard error of kurtosis, CI = Confidence interval.

Table 20. *Paired Comparison Results for Mean Length of AS Units for the Comparison Group*

	t	*df*	*p*	Cohen's *d*
Time 1-Time 2	-0.24	20	.810	0.09
Time 1-Time 3	-5.82	20	<.001	1.33
Time 2-Time 3	-3.84	20	.002	1.73

The next repeated-measures ANOVA was run with the data from the teacher-led group. Mauchly's test indicated that the assumption of sphericity was met, chi-square = 3.94, *p* = .14. The repeated-measures ANOVA indicated that there was no significant difference, *F*(2, 26) = 3.49, *p* = .045, partial η^2 = .21; thus, the participants in the teacher-led group did not significantly improve mean length of AS units over the treatment.

The third repeated-measures ANOVA was run with the data from the teacher and peer group. Mauchly's test indicated that the assumption of sphericity was met, chi-square = 3.82, *p* = .15. There was a significant difference, *F*(2, 40) = 9.48, *p* < .001, partial η^2 = .32, so three paired samples *t*-tests were run to make post hoc comparisons. The results are shown in Table 21. There was a significant decrease between Time 1 and Time 2, and a significant gain between Time 2 and Time 3.

Table 21. *Paired Comparison Results for Mean Length of AS Units for the Teacher and Peer Group*

	t	*df*	*p*	Cohen's *d*
Time 1-Time 2	4.18	20	< .001	0.96
Time 1-Time 3	-0.57	20	.58	0.12
Time 2-Time 3	-4.16	20	< .001	0.94

Figure 5 shows the changes in the mean length of AS units for the comparison group, the teacher-led group, and the teacher and peer group with transformed data. The comparison group significantly improved mean length of AS units from Time 1 to Time

3. The teacher-led group improved but not to a statistically significant degree, and the teacher and peer group decreased significantly from Time 1 to Time 2, but increased significantly from Time 2 to Time 3. However, there was no significant development of mean length of AS units over the academic semester.

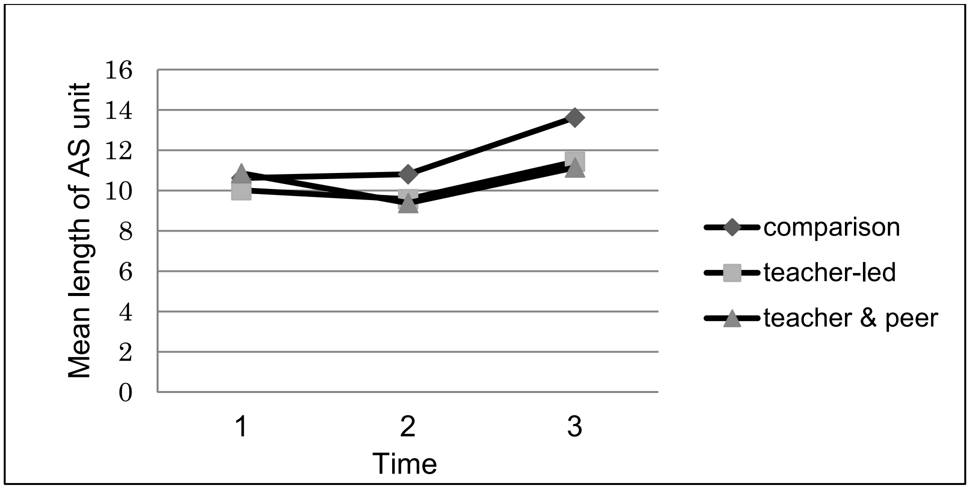


Figure 5. Changes in mean length of AS units at Times 1, 2, and 3.

Accuracy. Table 22 shows the descriptive statistics for error-free AS units for each group at Time 1, Time 2, and Time 3. The comparison group's percentage of error-free AS units decreased over the three times (Time 1: 71%, Time 2: 60%, Time 3: 59%), while the teacher-led group increased its percentage of error-free AS units over the three times (Time 1: 60%, Time 2: 64%, Time 3: 73%). The teacher and peer group changed only slightly throughout the semester (Time 1: 66%, Time 2: 62%, Time 3: 63%).

Table 22. *Descriptive Statistics for Error-Free AS Units at Times 1, 2, and 3*

	Comparison group			Teacher-led group			Teacher and peer group		
	Time 1	Time 2	Time 3	Time 1	Time 2	Time 3	Time 1	Time 2	Time 3
Minimum	0.38	0.18	0.36	0.33	0.20	0.40	0.44	0.11	0.27
Maximum	1.00	0.92	0.92	0.88	0.91	1.00	0.88	0.92	1.00
M	0.71	0.60	0.59	0.60	0.64	0.73	0.66	0.62	0.63
95% CI	[.62, .80]	[.49, .71]	[.49, .69]	[.51, .70]	[.53, .76]	[.64, .82]	[.60, .72]	[.51, .73]	[.55, .72]
SD	0.16	0.19	0.17	0.17	0.20	0.16	0.13	0.24	0.19
Skewness	-0.28	-0.67	0.44	0.33	-0.86	-0.26	-0.20	-0.58	0.05
SES	0.62	0.62	0.62	0.60	0.60	0.60	0.50	0.50	0.50
z-skewness	-0.45	-1.08	0.72	0.55	-1.45	-0.44	-0.40	-1.17	0.10
Kurtosis	0.99	1.33	-0.71	-0.96	0.71	0.27	-1.09	-0.63	-0.38
SEK	1.19	1.19	1.19	1.15	1.15	1.15	0.97	0.97	0.97
z-kurtosis	0.83	1.12	-0.60	-0.83	0.62	0.23	-1.12	-0.65	-0.39

Note. SES = Standard error of skewness, SEK = Standard error of kurtosis, CI = Confidence interval.

Before conducting the repeated-measures ANOVAs, the assumptions of the analyses were checked for each group at each of the three times (see Table 22). First, univariate outliers were checked by converting the raw scores for the speech data to *z*-scores and checking for values > ± 3.29 (Field, 2013). No univariate outliers were found. The assumption of normality was checked and met as shown by the *z*-skewness and *z*-kurtosis results.

Three repeated-measures ANOVAs were run with the dependent variable being error-free AS units. The first repeated-measures ANOVA was run with data from the comparison group. Mauchly's test indicated that the assumption of sphericity was met, chi-square = 3.55, $p = .17$. The results indicated a significant time effect for the comparison group, $F(2, 24) = 5.44$, $p = .011$, partial $\eta^2 = .31$. Three paired-samples *t*-tests were run as post hoc comparisons. As shown in Table 23, there was a significant decrease in accuracy between Time 1 and Time 2, and between Time 1 and Time 3. These results indicated that the accuracy of the comparison group participants decreased significantly at Time 2, and their accuracy remained low at Time 3.

Table 23. *Paired Comparison Results for Error-Free AS Units for the Comparison Group*

	t	*df*	*p*	Cohen's *d*
Time 1-Time 2	2.26	20	.043	0.63
Time 1-Time 3	3.64	20	.003	1.09
Time 2-Time 3	0.16	20	.880	0.08

The next repeated-measures ANOVA was run with the teacher-led group. Mauchly's test indicated that the assumption of sphericity was met, chi-square = .14, $p = .93$. There was no significant time effect, $F(2, 26) = 2.27$, $p = .12$, partial $\eta^2 = .15$.

The third repeated-measures ANOVA was run with the teacher and peer group. Mauchly's test indicated that the assumption of sphericity was met, chi-square = 3.47, *p* = .17. There was no significant time effect, *F*(2, 40) =.40, *p* = .67, partial η^2 = .020.

Figure 6 shows the changes in the percentage of error-free AS units for the comparison, the teacher-led, and the teacher and peer groups. The comparison group's accuracy scores decreased significantly, and there was no significant development of accuracy throughout the semester for the teacher-led group and the teacher and peer group. While the comparison group and the teacher and peer group's accuracy scores declined from Time 1 to Time 3, the descriptive statistics showed that the teacher-led group improved linearly. Thus, the pedagogical intervention had little effect on morphosyntactic accuracy.

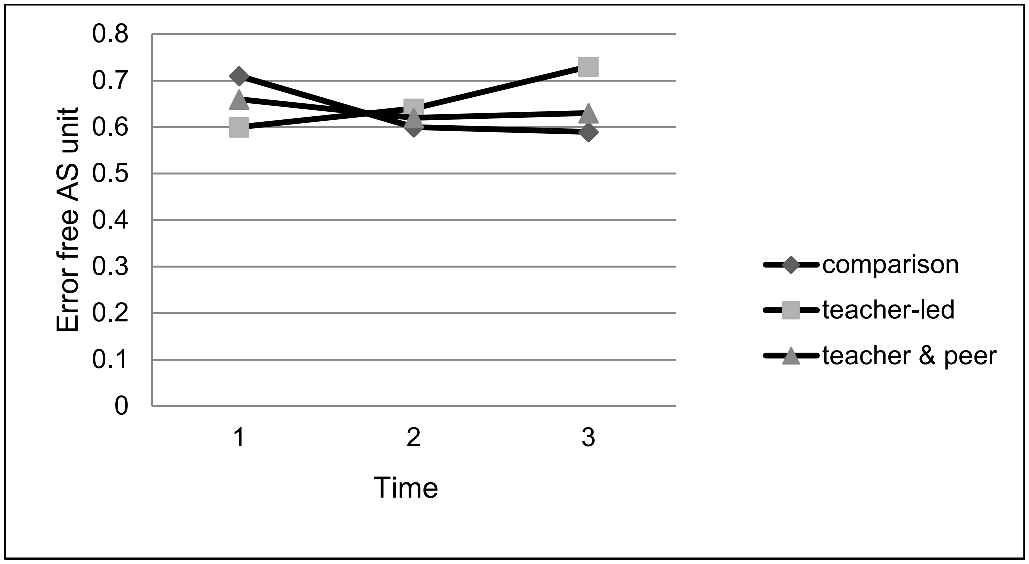


Figure 6. Changes in percentage of error-free AS units at Times 1, 2 and 3.

Lexis. MTLD was used as the measure of lexical diversity. As shown in Table 24, all three groups increased MTLD from Time 1 to Time 3, but they decreased slightly at Time 2.

Before conducting the repeated-measures ANOVAs, the assumptions of the analyses were checked for each group at each of the three times. First, univariate outliers were checked by converting the raw scores for the speech data to *z*-scores and checking for values > ± 3.29 (Field, 2013). No univariate outliers were found. Second, the assumption of normality was checked and met as shown by *z*-skewness and *z*-kurtosis (Table 24).

Three repeated-measures ANOVAs were run to investigate whether there was a significant development in lexical diversity in each group. Mauchly's test indicated that the assumption of sphericity was met for the comparison group, chi-square = .86, $p = .64$. There was a significant time effect, $F(2, 24) = 19.26$, $p < .001$, partial $\eta^2 = .62$. Three paired-samples *t*-tests were used to make post hoc comparisons. The results are displayed in Table 25. There was a significant gain between Time 1 and Time 3, and between Time 2 and Time 3. These results indicated that the comparison group participants significantly improved lexical diversity throughout the academic semester.

Table 24. *Descriptive Statistics for Measure of Textual Lexical Diversity at Times 1, 2, and 3*

	Comparison group			Teacher-led group			Teacher and peer group		
	Time 1	Time 2	Time 3	Time 1	Time 2	Time 3	Time 1	Time 2	Time 3
Minimum	23.78	23.48	29.26	25.69	24.33	29.93	20.32	18.19	22.69
Maximum	50.65	36.81	61.53	52.19	50.07	61.58	59.49	46.22	64.03
M	34.47	30.03	46.26	37.62	33.60	44.06	37.38	30.04	41.82
95% CI	[29.45, 39.50]	[27.16, 32.90]	[40.26, 52.39]	[32.33, 42.91]	[29.33, 37.87]	[38.51, 49.61]	[32.58, 42.19]	[26.69, 33.39]	[36.69, 46.95]
SD	8.32	4.75	10.16	9.16	7.40	9.61	10.56	7.36	11.27
Skewness	0.27	0.11	0.14	0.47	0.84	0.05	0.20	0.63	0.53
SES	0.62	0.62	0.62	0.60	0.60	0.60	0.50	0.50	0.50
z-skewness	0.44	0.18	0.22	0.78	1.41	0.08	0.40	1.25	1.06
Kurtosis	-0.79	-1.59	-1.08	-1.24	0.05	-0.91	-0.61	0.20	0.09
SEK	1.19	1.19	1.19	1.15	1.15	1.15	0.97	0.97	0.97
z-kurtosis	-0.66	-1.33	-0.91	-1.08	0.05	-0.79	-0.63	0.21	0.10

Note. SES = Standard error of skewness, SEK = Standard error of kurtosis, CI = Confidence interval.

Table 25. *Paired Comparison Results for Measure of Textual Lexical Diversity for the Comparison Group*

	t	*df*	*p*	Cohen's *d*
Time 1-Time 2	1.93	20	.080	0.57
Time 1-Time 3	-4.05	20	.002	1.13
Time 2-Time 3	-5.69	20	< .001	1.72

The next repeated-measures ANOVA was run with the teacher-led group. Mauchly's test indicated that the assumption of sphericity was met, chi-square = 5.50, p = .06. There was a significant time effect, $F(2, 26) = 6.58$, $p = .005$, partial $\eta^2 = .34$, so three paired samples *t*-tests were used to make post hoc comparisons. There was a significant gain between Time 2 and Time 3 (Table 26).

Table 26. *Paired Comparison Results for Measure of Textual Lexical Diversity for the Teacher-Led Group*

	t	*df*	*p*	Cohen's *d*
Time 1-Time 2	1.29	20	.220	0.35
Time 1-Time 3	-1.86	20	.085	0.50
Time 2-Time 3	-5.62	20	< .001	1.57

The third repeated-measures ANOVA was run with the teacher and peer group. Mauchly's test indicated that the assumption of sphericity was met, chi-square = .68, p = .71. There was a significant difference, $F(2, 40) = 9.76$, $p < .001$, partial $\eta^2 = .33$, so three paired-samples *t*-tests were used to make post hoc comparisons among Time 1, Time 2, and Time 3. There was a significant decline from Time 1 to Time 2, and a significant gain between Time 2 and Time 3 (Table 27).

Table 27. *Paired Comparison Results for Measure of Textual Lexical Diversity for the Teacher and Peer Group*

	t	*df*	*p*	Cohen's *d*
Time 1-Time 2	2.98	20	.007	0.67
Time 1-Time 3	-1.53	20	.140	0.33
Time 2-Time 3	-4.37	20	< .001	1.02

Figure 7 shows the changes in MTLD for the comparison, teacher-led, and teacher and peer groups at Time 1, Time 2, and Time 3. All three groups displayed significant development from Time 2 to Time 3, in part because the groups did not perform well at Time 2. The topic at Time 2 was *eating out*. Given that students might not need to use a wide variety of vocabulary to discuss eating compared to topics such as club activity (Time 1) or English learning (Time 2), the decrease in lexical diversity might have been due to the topic. However, when comparing Time 1 and Time 3, only the comparison group showed a significant development.

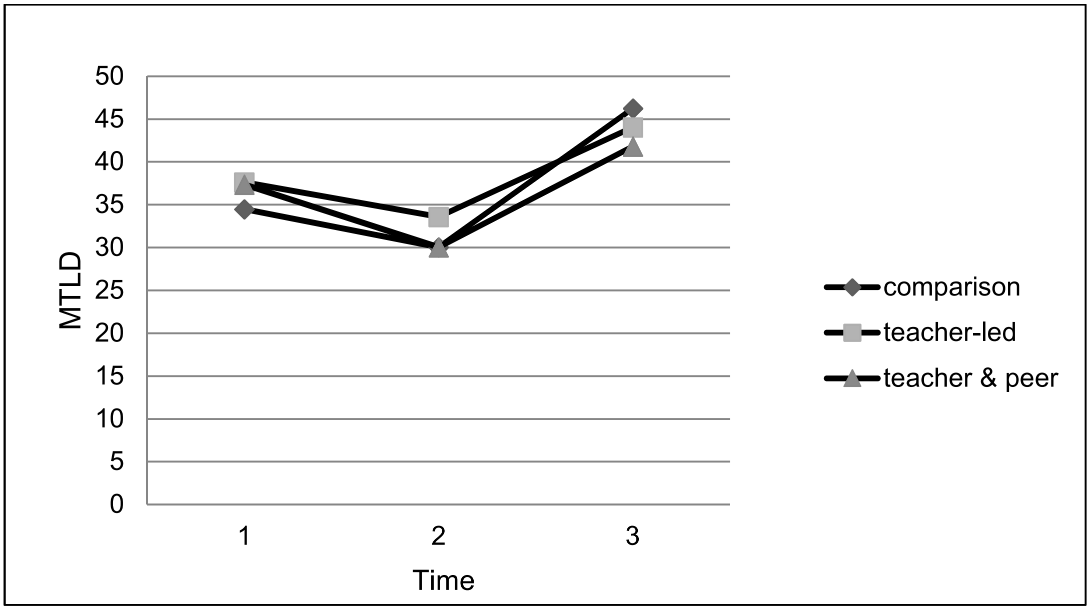


Figure 7. Changes in Measure of Textual Lexical Diversity at Times 1, 2, and 3.

Mean length of pauses (fluency). Table 28 shows the descriptive statistics for mean length of pauses at Time 1, Time 2, and Time 3. Mean length of pauses shows breakdown fluency; larger numbers indicate more disfluency. The comparison group did not change mean length of pause over the academic semester, while the teacher-led group decreased slightly over the three test occasions (Time 1 = 1.03, Time 2 = 1.03, Time 3 = 1.01), and the teacher and peer group increased slightly throughout the semester (Time 1 = 0.99, Time 2 = 0.99, Time 3 = 1.04).

Before conducting the repeated-measures ANOVAs, the assumptions of the analyses were checked for each group at each of the three times. First, univariate outliers were checked by converting the raw scores for the speech data to *z*-scores and checking for values > ± 3.29 (Field, 2013). No univariate outliers were found. Second, the assumption of normality was checked and met as shown by the *z*-skewness and *z*-kurtosis statistics (Table 28).

Table 28. *Descriptive Statistics for Mean Length of Pauses at Time 1, 2, and 3*

	Comparison group			Teacher-led group			Teacher and peer group		
	Time 1	Time 2	Time 3	Time 1	Time 2	Time 3	Time 1	Time 2	Time 3
Minimum	.44	.62	.84	.43	.33	1.00	.30	.37	.87
Maximum	1.25	1.07	1.32	1.05	1.14	.88	.77	.98	2.58
M	.75	.86	1.05	.71	.72	2.29	.59	.62	1.57
95% CI	[.64, .87]	[.79, .93]	[.96, 1.13]	[.62, .81]	[.62, .83]	[1.13, 1.50]	[.53, .65]	[.55, .69]	[1.40, 1.78]
SD	.21	.13	.15	.18	.20	.38	.13	.17	.43
Skewness	.96	-.41	.56	.38	.11	1.83	-.46	.41	.73
SES	.62	.62	.62	.60	.60	.60	.50	.50	.50
z-skewness	1.55	-.66	.90	.63	.18	3.05	-.92	.82	1.46
Kurtosis	1.81	-.04	-.86	-.75	.59	3.28	-.20	-.63	.19
SEK	1.19	1.19	1.19	1.15	1.15	1.15	.97	.97	.97
z-kurtosis	1.52	-.03	-.72	-.65	.51	2.85	-.21	-.65	.20

Note. SES = Std. Error Skewness, SEK = Std. Error kurtosis, CI = Confidence interval.

Three repeated-measures ANOVAs were run to investigate whether there was significant time effect in mean length of pauses in each group. The first repeated-measures ANOVA was run with data from the comparison group. Mauchly's test indicated that the assumption of sphericity was met, chi-square = 1.18, $p = .55$. There was a significant time effect $F(2, 24) = 14.54$, $p < .001$, partial $\eta^2 = .55$, so three paired-samples *t*-tests were run to make post hoc comparisons. The results are shown in Table 29. There was a significant increase in pauses between Time 1 and Time 3, and Time 2 and Time 3. Thus, the participants in the comparison group increased mean length of pauses significantly throughout the academic semester.

Table 29. *Paired Comparison Results for Mean Length of Pauses for the Comparison Group*

	t	*df*	*p*	Cohen's *d*
Time 1-Time 2	-2.47	20	.029	0.71
Time 1-Time 3	-4.89	20	< .001	1.68
Time 2-Time 3	-3.11	20	.009	1.31

Mauchly's test indicated that the assumption of sphericity was violated, chi-square = 15.09, $p = .001$; therefore, the Greenhouse-Geisser results are reported. The next repeated-measures ANOVA was run with the teacher-led group. There was a significant time effect, $F(2, 26) = 18.62$, $p < .001$, partial $\eta^2 = .59$, so three paired samples *t*-tests were run to make post hoc comparisons among Time 1, Time 2, and Time 3. The results are shown in Table 30. There was a significant increase between Time 1 and Time 3, and between Time 2 and Time 3. Thus, the participants in the teacher-led group significantly increased pause length at Time 3.

Table 30. *Paired Comparison Results for Mean Length of Pauses for the Teacher-Led Group*

	t	*df*	*p*	Cohen's *d*
Time 1-Time 2	-0.30	20	.770	0.05
Time 1-Time 3	-4.48	20	.001	2.14
Time 2-Time 3	-4.38	20	.001	2.03

The next repeated-measures ANOVA was run with the teacher and peer group. The assumption of sphericity was violated, chi-square = 26.27, $p < .001$; therefore, the Greenhouse-Geisser results are reported. There was a significant time effect, $F(2, 40) = 68.71$, $p < .001$, partial $\eta^2 = .78$, so three paired-samples *t*-tests were used to make post hoc comparisons between conditions (see Table 31). There was a significant increase between Time 1 and Time 3, and between Time 2 and Time 3; thus, the participants in the teacher and peer group significantly increased mean length of pauses throughout the academic semester.

Table 31. *Paired Comparison Results for Mean Length of Pauses for the Teacher and Peer Group*

	t	*df*	*p*	Cohen's *d*
Time 1-Time 2	-0.83	20	.410	0.20
Time 1-Time 3	-9.40	20	.001	3.54
Time 2-Time 3	-7.82	20	< .001	3.20

Figure 8 shows the changes in mean length of pauses for the comparison, the teacher-led, the teacher and peer group at Time 1, Time 2, and Time 3. The participants in all three groups increased the mean length of pauses at Time 3.

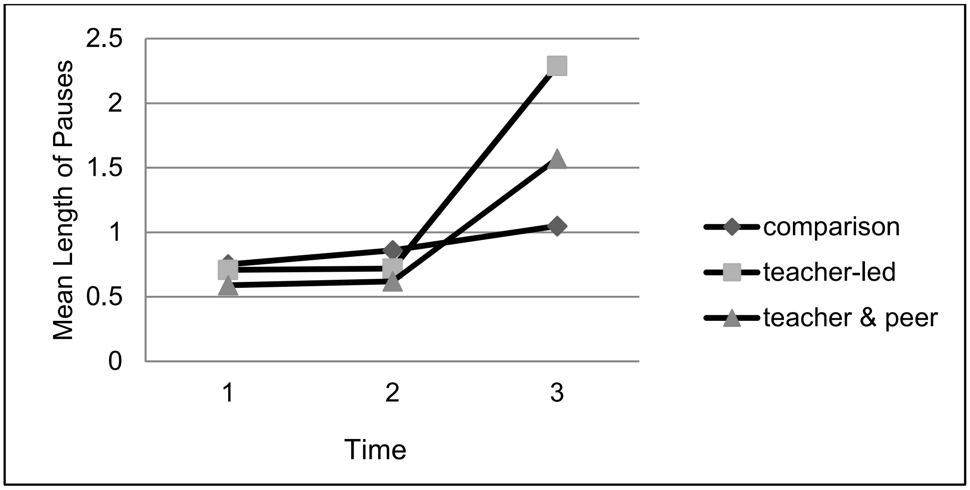


Figure 8. Changes in mean length of pauses at Time 1, 2, and 3.

Repairs (fluency). Table 32 shows the descriptive statistics for repairs at Time 1, Time 2, and Time 3. Number of repairs is a measure of repair fluency, so a larger number indicates disfluency. The comparison group had similar results at the three times: Time 1 = 12.15, Time 2 = 12.31, Time 3 = 11.69. The teacher-led group produced the same number of repairs at Time 1 and Time 3. On the other hand, the teacher and peer group produced fewer repairs than the comparison group and the teacher-led group. Throughout the semester, the teacher and peer group had 7-8 repair occurrences, showing that the group did not produce many repairs from the beginning of the study. The standard deviation and skewness statistics, however, indicate that the repair occurrences varied depending on the individual speaker. For example, at Time 1, one person in the comparison group produced only two repairs while another person produced 35 repairs. The maximum number of occurrences was 30, while the minimum was 3 In the teacher-led group at Time 2.

Table 32. *Descriptive Statistics for Repairs at Time 1, 2, and 3.*

	Comparison group			Teacher-led group			Teacher and peer group		
	Time 1	Time 2	Time 3	Time 1	Time 2	Time 3	Time 1	Time 2	Time 3
Minimum	2.00	6.00	8.00	1.00	3.00	1.00	2.00	1.00	.00
Maximum	35.00	19.00	16.00	21.00	30.00	22.00	23.00	18.00	19.00
M	12.15	12.31	11.69	10.50	12.07	10.64	7.71	7.00	8.05
95% CI	[6.67, 17.63]	[9.47, 15.14]	[9.92, 13.46]	[7.10, 13.90]	[7.49, 16.65]	[7.34, 13.95]	[5.42, 10.00]	[4.94, 9.06]	[5.76, 10.33]
SD	9.07	4.70	2.93	5.88	7.93	5.72	5.03	4.52	5.01
Skewness	1.54	.14	.15	.55	.94	.16	1.64	.96	.71
SES	.62	.62	.62	.60	.60	.60	.50	.50	.50
z-skewness	2.50	.22	.24	.92	1.58	.27	3.27	1.92	1.42
Kurtosis	2.26	-1.48	-1.48	-.05	.43	-.12	3.26	.37	.23
SEK	1.19	1.19	1.19	1.15	1.15	1.15	.97	.97	.97
z-kurtosis	1.90	-1.25	-1.24	-.04	.37	-.10	3.36	.38	.23

Note. SES = Standard error of skewness, SEK = Standard error of kurtosis, CI = Confidence interval.

Before conducting the repeated-measures ANOVAs, the assumptions of the analyses were checked for each group at each of the three times. First, univariate outliers were checked by converting the raw scores for the speech data to *z*-scores and checking for values > ± 3.29 (Field, 2013). No univariate outliers were found. Second, the assumption of normality was checked and met as shown by *z*-skewness and *z*-kurtosis (Table 32).

Three repeated-measures ANOVAs were run to investigate whether there were significant changes in repairs for the three groups. The first repeated-measures ANOVA was run with data from the comparison group. Mauchly's test indicated that the assumption of sphericity was met, chi-square = 3.37, $p = .17$. There was no significant differences for the comparison group, $F(2, 24) = .05$, $p = .95$, partial $\eta^2 = .004$.

The next repeated-measures ANOVA was run with the teacher-led group. Mauchly's test indicated that the assumption of sphericity was met, chi-square = .71, $p = .70$. There was no significant difference for the teacher-led group, $F(2, 26) = .40$, $p = .67$, partial $\eta^2 = .03$.

The third repeated-measures ANOVA was run with the comparison group. Mauchly's test indicated that the assumption of sphericity was met, chi-square = 2.94, $p = .23$. There was no significant difference, $F(2, 40) = .77$, $p = .47$, partial $\eta^2 = .037$.

Figure 9 shows changes in repair for the comparison, the teacher-led, the teacher and peer group at Times 1, Time 2, and Time 3. The pedagogical treatment did not have an effect on repair fluency.

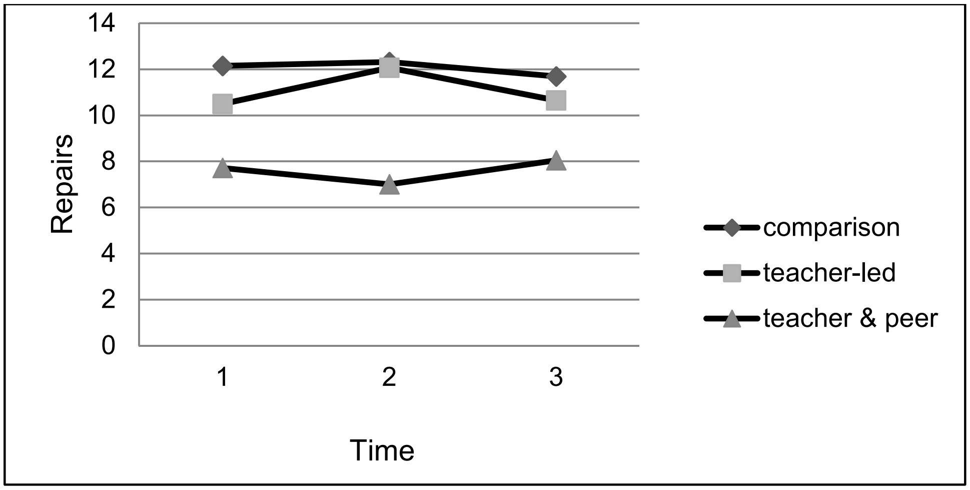


Figure 9. Changes in repairs at Times 1, 2 and 3.

Mean Duration of Syllable (Articulation rate). Table 33 shows the descriptive statistics for mean duration of syllable at Time 1, Time 2, and Time 3. Mean duration of syllable shows speed fluency. Longer syllable durations indicate disfluency. Because providing one syllable should take less than 1 second, the descriptive statistics show that the differences between each time were small. The mean duration of syllable of the comparison group ranged between 0.28 to 0.32, the teacher-led group ranged between 0.29 and 0.32, and the teacher and peer group ranged between 0.30 and 0.33.

Before conducting the repeated-measures ANOVAs, the assumptions of the analyses were checked for each group at each of the three times. First, univariate outliers were checked by converting the raw scores for the speech data to *z*-scores and checking for values > ± 3.29 (Field, 2013). No univariate outliers were found. Second, assumption of normality was checked and met as shown by *z*-skewness and *z*-kurtosis (Table 33).

Table 33. *Descriptive Statistics for Mean Duration of Syllable at Time 1, 2, and 3*

	Comparison group			Teacher-led group			Teacher and peer group		
	Time 1	Time 2	Time 3	Time 1	Time 2	Time 3	Time 1	Time 2	Time 3
Minimum	.22	.27	.25	.22	.28	.23	.23	.28	.26
Maximum	.33	.39	.37	.37	.37	.34	.42	.42	.40
M	.28	.32	.30	.29	.32	.29	.30	.33	.31
95% CI	[.26, .29]	[.29, .33]	[.29, .32]	[.26, .31]	[.30, .33]	[.27, .31]	[.28, .33]	[.31, .35]	[.29, .32]
SD	.03	.04	.03	.04	.03	.03	.05	.04	.03
Skewness	-.13	.76	.51	.36	.69	.02	.74	.44	1.37
SES	.62	.62	.62	.60	.60	.60	.50	.50	.50
z-skewness	-.20	1.23	.82	.61	1.16	.03	1.47	.88	2.73
Kurtosis	1.23	-.25	.71	-.44	-.79	-1.33	.26	-.49	3.54
SEK	1.19	1.19	1.19	1.15	1.15	1.15	.97	.97	.97
z-kurtosis	1.03	-.21	.59	-.38	-.68	-1.15	.27	-.50	3.65

Note. SES = Standard error of skewness, SEK = Standard error of kurtosis, CI = Confidence interval.

Three repeated-measures ANOVA were run to investigate whether there were significant changes in mean duration of syllable for the three groups. The first repeated-measures ANOVA was run with data from the comparison group. Mauchly's test indicated that the assumption of sphericity was met, chi-square = 1.52, *p* = .46. There was a significant time effect, *F*(2, 24) = 9.27, *p* = .001, partial η^2 = .44, so three paired samples *t*-tests were run to make post hoc comparisons. The results are shown in Table 34. There was a significant increase between Time 1 and Time 2, and between Time 1 and Time 3. This result shows that the comparison group's production of syllables became longer throughout the semester.

Table 34. *Paired Comparison Results for Mean Duration of Syllable for the Comparison Group*

	t	*df*	*p*	Cohen's *d*
Time 1-Time 2	-3.62	20	.004	1.11
Time 1-Time 3	-3.13	20	.009	0.66
Time 2-Time 3	1.53	20	.150	0.75

The next repeated-measures ANOVA was run with the teacher-led group. Mauchly's test indicated that the assumption of sphericity was met, chi-square = 1.21, *p* = .54. There was a significant time effect, *F*(2, 26) = 9.55, *p* = .001, partial η^2 = .42, so post-hoc paired-samples *t*-tests were run. The results are shown in Table 35. There was a significant increase between Time 1 and Time 2, and a significant decrease between Time 2 and Time 3.

Table 35. *Paired Comparison Results for Mean Duration of Syllable for the Teacher-Led Group*

	t	*df*	*p*	Cohen's *d*
Time 1-Time 2	-3.91	20	.002	1.31
Time 1-Time 3	0.16	20	.880	0.00
Time 2-Time 3	4.31	20	.001	1.20

The last repeated-measures ANOVA was run with the teacher and peer group. The assumption of sphericity, which refers to the equality of variances of the differences between treatment levels (Field, 2013, p. 459), was checked by running Mauchly's sphericity test. Mauchly's test indicated that the assumption of sphericity was violated, chi-square = 8.30, p = .016, so the Greenhouse-Geisser results are reported.

There was a significant time effect, $F(2, 40) = 5.63$, $p = .01$, partial $\eta^2 = .22$, so three paired-samples *t*-tests were run. The results are shown in Table 36. There was a significant increase between Time 1 and Time 2, but there was a significant decrease between Time 2 and Time 3; therefore, the participants in the teacher and peer group did not change mean length of pauses significantly between Time 1 and Time 3.

Table 36. *Paired Comparison Results for Mean Duration of Syllable for the Teacher and Peer Group*

	t	*df*	*p*	Cohen's *d*
Time 1-Time 2	-2.99	20	.007	0.63
Time 1-Time 3	-0.27	20	.790	0.18
Time 2-Time 3	4.00	20	.001	0.64

Figure 10 shows the changes in mean duration of syllable for the comparison group, the teacher-led group, and the teacher and peer group at Time 1, Time 2, and Time 3. All groups displayed similar changes throughout the semester; they had the lowest mean score at Time 1, they significantly increased mean duration of syllable at Time 2, and they decreased at Time 3. For the teacher-led and the teacher and peer group, there was a statistically significant decline between Time 2 and Time 3. Because the longer mean duration of syllable indicates more disfluency, the participants' fluency improved from Time 2 to Time 3, but it did not show significant improvement throughout the treatment.

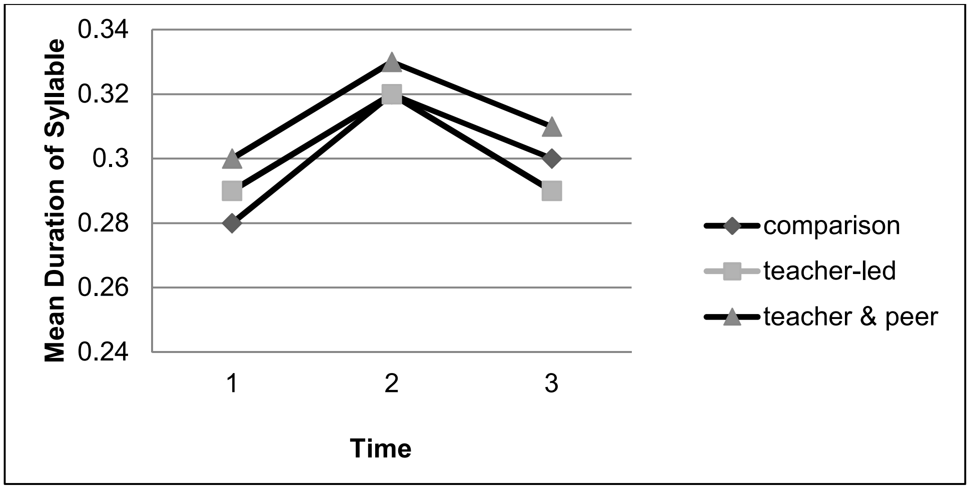


Figure 10. Changes in mean duration of syllable at Times 1, 2, and 3.

Mean length of run (fluency). Table 37 shows the descriptive statistics for mean length of run at Time 1, Time 2, and Time 3. Mean length of run is a combined measure of both breakdown and speed fluency. Higher scores indicate greater fluency. The comparison group had similar scores at Time 1 (5.73) and Time 2 (5.87) and they increased slightly at Time 3 to 6.10. The teacher-led group increased from Time 1 (4.58) to Time 3 (5.31), although they decreased slightly at Time 2 (4.33). The teacher and peer group increased mean length of run from Time 1 (3.95) to Time 2 (4.50) and to Time 3 (5.14).

Table 37. *Descriptive Statistics for Mean Length of Run at Times 1, 2, and 3*

	Comparison group			Teacher-led group			Teacher and peer group		
	Time 1	Time 2	Time 3	Time 1	Time 2	Time 3	Time 1	Time 2	Time 3
Minimum	3.70	3.10	3.65	3.21	2.83	4.44	2.61	3.00	3.50
Maximum	5.73	5.87	6.10	5.95	5.31	7.00	6.32	7.21	7.23
M	4.88	4.34	5.05	4.58	4.33	5.31	3.95	4.50	5.14
95% CI	[4.42, 5.53]	[3.87, 4.80]	[4.57, 5.53]	[4.16, 5.00]	[3.83, 4.81]	[4.81, 5.80]	[3.60, 4.30]	[3.96, 5.04]	[4.71, 5.56]
SD	.75	.78	.80	.72	.84	.86	.76	1.18	.93
Skewness	-.34	.33	-.04	.21	-.41	1.07	1.30	.90	.43
SES	.62	.62	.62	.60	.60	.60	.50	.50	.50
z-skewness	-.56	.53	-.06	.36	-.69	1.80	2.59	1.80	.86
Kurtosis	-1.30	-.28	-1.06	.25	-1.32	.04	3.96	.22	.40
SEK	1.19	1.19	1.19	1.15	1.15	1.15	.97	.97	.97
z-kurtosis	-1.09	-.23	-.89	.22	-1.14	.03	4.08	.22	.41

Note. SES = Standard error of skewness, SEK = Standard error of kurtosis, CI = Confidence interval.

Before conducting the repeated-measures ANOVAs, the assumptions of the analyses were checked for each group at each of the three times. First, univariate outliers were checked by converting the raw scores for the speech data to *z*-scores and checking for values > ± 3.29 (Field, 2013). No univariate outliers were found. Second, the assumption of normality was checked and met as shown by the *z*-skewness and *z*-kurtosis statistics (Table 37).

Three repeated-measures ANOVA were run to investigate whether there was a significant effect for mean length of run for the three groups. The first repeated-measures ANOVA was run with data from the comparison group. Mauchly's test indicated that the assumption of sphericity was met, chi-square = .14, p = .93. There was no significant time effect, F(2, 24) = 5.07, p = .015, partial η^2 = .29; thus, the comparison group did not improve mean length of run over the semester.

The next repeated-measures ANOVA was run with the teacher-led group. Mauchly's test indicated that the assumption of sphericity was met, chi-square = 1.07, p = .57. There was a significant time effect, F(2, 26) = 10.00, p = .001, partial η^2 = .44, so three paired-samples *t*-tests were run to make post hoc comparisons. The results are shown in Table 38. There was significant increase between Time 1 and Time 3, and between Time 2 and Time 3.

Table 38. *Paired Comparison Results for Mean Length of Run for the Teacher-Led Group*

	t	*df*	*p*	Cohen's *d*
Time 1-Time 2	1.30	20	.220	0.35
Time 1-Time 3	-2.96	20	.010	0.79
Time 2-Time 3	-4.07	20	.001	1.08

The third repeated-measures ANOVA was run with the comparison group. Mauchly's test indicated that the assumption of sphericity was met, chi-square = .45, p = .80. There was a significant time effect for the teacher and peer group, $F(2, 40) = 17.63$, $p < .001$, partial $\eta^2 = .44$, so three paired-samples *t*-tests were run. The results are shown in Table 39. There was a significant increase between Time 1 and Time 3, and between Time 2 and Time 3.

Table 39. *Paired Comparison Results for Mean Length of Run for the Teacher and Peer Group*

	t	*df*	*p*	Cohen's *d*
Time 1-Time 2	-2.57	20	.018	0.61
Time 1-Time 3	-5.98	20	< .001	1.31
Time 2-Time 3	-3.41	20	.003	0.77

Figure 11 shows the changes in mean length of run for the comparison group, the teacher-led group, the teacher and peer group at Time 1, Time 2, and Time 3. The comparison group declined from Time 1 to Time 2. Because of the large decline, there was a statistically significant change from Time 2 to Time 3. The teacher-led group also declined at Time 2 from Time 1, however, the group made a statistically significant improvement at Time 3. Both the teacher-led group and the teacher and peer group significantly improved throughout the semester.

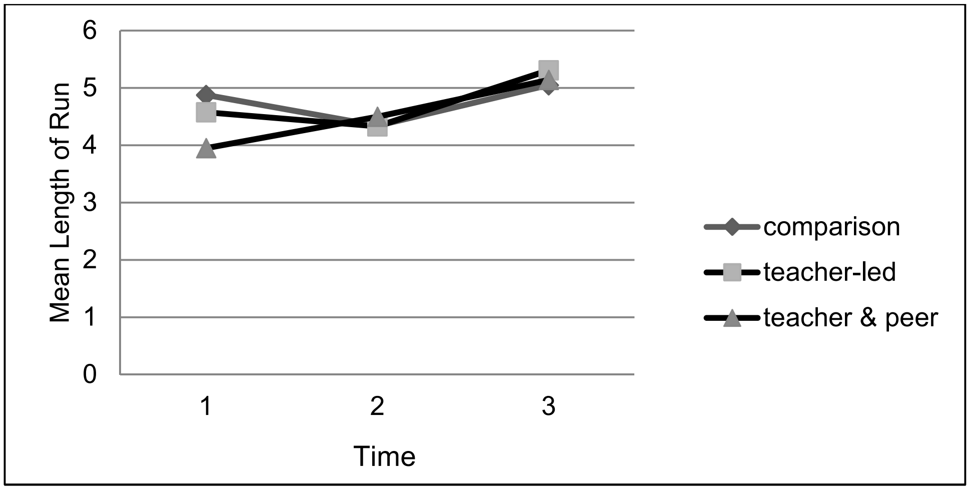


Figure 11. Changes in mean length of run at Times 1, 2, and 3.

A further analysis was conducted to more fully investigate mean length of run. A descriptive analysis was conducted in order to examine which group improved MLR the most. The MLR gain scores (MLR at Time 3 – MLR at Time 2 = MLR gain) were converted into *z*-scores as shown in Table 40. The top six students' *z* scores were more than 1 standard deviation above the mean: Student 1 (*z* = 2.67), Student 2 (*z* = 1.86), Student 3 (*z* = 1.80), Student 4 (*z* = 1.61), Student 5 (*z* = 1.42), and Student 6 (*z* = 1.36) (See Table 40). Five of these students were from the teacher and peer group.

Table 40. *Mean Length of Run Gains and Usage of the Target Form*

Student	Group	MRL gains	Target form frequency (Time 1)	Target form frequency (Time 3)	Frequency gain (Time 3 - Time 1)	*z*-score
1	Teacher & peer	+3.41	3	8	+5	2.68
2	Teacher-led	+2.61	1	1	0	1.86
3	Teacher & peer	+2.54	1	6	+5	1.80
4	Teacher & peer	+2.36	3	9	+6	1.61
5	Teacher & peer	+2.17	1	7	+6	1.42
6	Teacher & peer	+2.11	1	8	+7	1.36

Furthermore, these participants used more target formulaic language at Time 3 compared to Time 1. This finding suggests that mean length of run gains might be associated with the usage of the target formulaic language. This issue is discussed in Chapter 5.

Phonation time ratio (fluency). Table 41 shows the descriptive statistics for phonation time ratio at Time 1, Time 2, and Time 3. Phonation time ratio is a combined measure of oral fluency, as both breakdowns and speed are part of this measure. This measure was calculated as the percentage of speaking time in the total time. Higher percentages indicate greater fluency. Overall, all groups increased the phonation time ratio. For example, the comparison group increased from 54% (Time 1) to 61% (Time 3), the teacher-led group increased from 52% (Time 1) to 55% (Time 3), and the teacher and peer group increased from 45% (Time 1) to 51% (Time 3). The teacher-led group and the teacher and peer group spent approximately half of the task speaking; thus, they spoke for approximately 1 minute in the 2-minute task.

Before conducting the repeated-measures ANOVAs, the assumptions of the analyses were checked for each group at each of the three times. First, univariate outliers were checked by converting the raw scores for the speech data to *z*-scores and checking for values > ± 3.29 (Field, 2013). No univariate outliers were found. Second, the assumption of normality was checked and met as shown by the *z*-skewness and *z*-kurtosis statistics (Table 41).

Table 41. *Descriptive Statistics for Phonation Time Ratio at Times 1, 2, and 3*

	Comparison group			Teacher-led group			Teacher and peer group		
	Time 1	Time 2	Time 3	Time 1	Time 2	Time 3	Time 1	Time 2	Time 3
Minimum	41.46	44.62	52.38	36.08	34.92	41.45	35.46	32.90	35.45
Maximum	69.85	65.55	65.79	67.75	64.46	68.85	56.79	70.70	67.98
M	53.97	55.19	60.83	52.31	49.68	55.61	45.50	48.48	51.85
95% CI	[48.82, 59.11]	[51.63, 58.76]	[58.22, 63.45]	[46.74, 57.88]	[44.14, 55.22]	[51.08, 60.14]	[42.72, 48.28]	[43.72, 53.23]	[47.80, 55.89]
SD	8.51	5.90	4.33	9.65	9.59	7.85	6.11	10.44	8.88
Skewness	.37	.30	-.68	-.25	-.24	-.26	.26	.55	.04
SES	.62	.62	.62	.60	.60	.60	.50	.50	.50
z-skewness	.61	.48	-1.10	-.41	-.40	-.44	.52	1.10	.07
Kurtosis	-.10	.18	-.65	-.59	-.85	-.30	-.46	-.61	-.48
SEK	1.19	1.19	1.19	1.15	1.15	1.15	.97	.97	.97
z-kurtosis	-.09	.15	-.55	-.51	-.73	-.26	-.47	-.63	-.49

Note. SES = Standard error of skewness, SEK = Standard error of kurtosis, CI = Confidence interval.

Three repeated-measures ANOVAs were run to investigate whether there was significant growth in phonation time ratio for the three groups. The first repeated-measures ANOVA was run with data from the comparison group. Mauchly's test indicated that the assumption of sphericity was met, chi-square = 3.46, *p* = .18. There were a significant difference, *F*(2, 24) = 7.01, *p* = .004, partial η^2 = .37, so three paired-samples *t*-tests were run to make post hoc comparisons. The results are shown in Table 42. There was no significant difference between Time 1 and Time 2 nor between Time 1 and Time 3; however, there was a significant increase between Time 2 and Time 3. This finding indicates that the participants in the comparison group did not develop significantly in the first half of the semester, but they significantly improved phonation time ratio in the latter part of the semester.

Table 42. *Paired Comparison Results for Phonation Time Ratio for the Comparison Group*

	t	*df*	*p*	Cohen's *d*
Time 1-Time 2	-0.60	20	.560	0.17
Time 1-Time 3	-2.97	20	.012	0.90
Time 2-Time 3	-4.02	20	.002	1.16

The next repeated-measures ANOVA was run with the teacher-led group. Mauchly's test indicated that the assumption of sphericity was met, chi-square = .34 *p* = .84. A repeated-measures ANOVA showed that there was no significant time effect, *F*(2, 26) = 4.40, *p* = .023, partial η^2 = .25. The participants in the teacher-led group did not significantly change phonation time ratio throughout the semester.

The last repeated-measures ANOVA was run with the teacher and peer group. Mauchly's test indicated that the assumption of sphericity was met, chi-square = .41, *p* =

.82. A repeated-measures ANOVA indicated that there was a significant time effect, *F*(2, 40) = 5.86, *p* = .006, partial η^2 = .23. Three paired-samples *t*-tests were used to make post hoc comparisons (Table 43). No significant difference was found between Time 1 and Time 2 or between Time 2 and Time 3; however, there was a significant gain between Time 1 and Time 3.

Table 43. *Paired Comparison Results for Phonation Time Ratio for the Teacher and Peer Group*

	t	*df*	*p*	Cohen's *d*
Time 1-Time 2	-1.54	20	.140	0.37
Time 1-Time 3	-3.33	20	.003	0.76
Time 2-Time 3	-1.97	20	.063	0.44

Figure 12 shows changes in phonation time ratio for the comparison group, the teacher-led group, and the teacher and peer group at Time 1, Time 2, and Time 3.

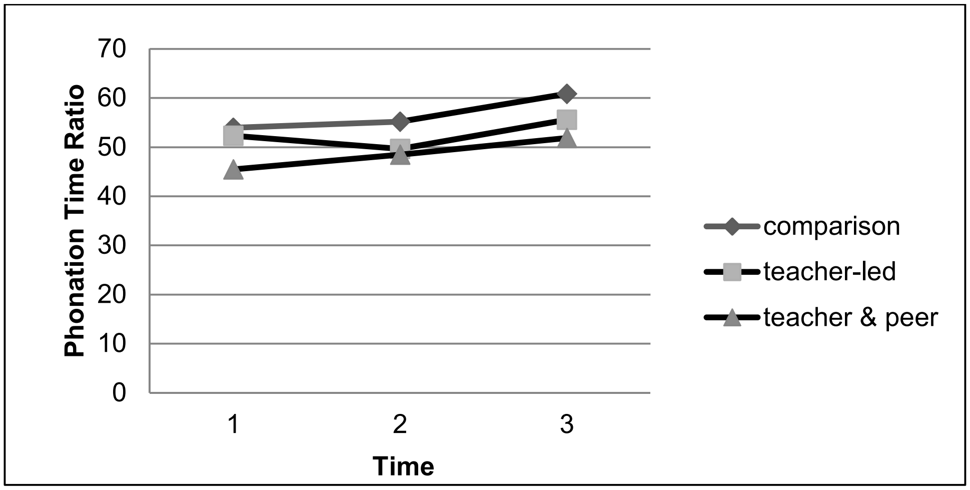


Figure 12. Changes in phonation time ratio at Time 1, 2 and 3.

The comparison group improved significantly from Time 2 to Time 3, and the teacher and peer group improved significantly from Time 1 to Time 3. On the other hand, the teacher-led group declined at Time 2.

Summarizing the results for research question 1, the comparison group increased syntactic complexity significantly between Time 1 and Time 3 for clauses per AS unit (*p* = .001, Cohen's *d* = 1.27) and mean length of AS units (*p* < .001, Cohen's *d* = 1.33). They also significantly increased lexical diversity between Time 1 and Time 3 (*p* = .002, Cohen's *d* = 1.13). They significantly increased mean duration of syllable between Time 1 and Time 2 (*p* = .004, Cohen's *d* = 1.11) and between Time 1 and Time 3 (*p* = .009, Cohen's *d* = .66), mean length of pauses between Time 1 and Time 3 (*p* < .001, Cohen's *d* = 1.68), which implies more disfluent. On the other hand, they significantly increased phonation time ratio between Time 2 and Time 3 (*p* = .002, Cohen's *d* = 1.16).

The teacher-led group significantly increased clauses per AS unit between Time 1 and Time 3, *p* = .01, Cohen's *d* = .87. They significantly increased mean length of pauses between Time 1 and Time 3 (*p* = .001, Cohen's *d* = 2.14) and between Time 2 and Time 3 (*p* = .001, Cohen's *d* = 2.03). However, they significantly increased mean length of run between Time 1 and 3 (*p* = .01, Cohen's *d* = .79) and Time 2 and Time 3 (*p* = .001, Cohen's *d* = 1.08).

The teacher and peer group significantly increased mean length of pauses between Time 1 and Time 3 (*p* = .001, Cohen's *d* = 3.54) and between Time 2 and 3 (*p* < .001, Cohen's *d* = 3.20). However, they significantly increased mean length of run between Time 1 and Time 3 (*p* < .001, Cohen's *d* = 1.31) and between Time 2 and Time 3 (*p* =

.003, Cohen's $d = .77$), and phonation time ratio between Time 1 and Time 3 ($p = .003$, Cohen's $d = .76$).

Table 44 shows a summary of the changes in the CALF measures throughout the semester. Upward arrows indicate an increase in the measure and the downward arrows indicate a decline between Time 1 and Time 3. The comparison group significantly gained complexity and lexical diversity, but they decreased syntactic accuracy (error free

Table 44. *Summary of CALF Development*

Measure	Time	Comparison group	Teacher-led group	Teacher and peer group
Clauses per AS-unit	1-2	*ns*	*ns*	$p = .010$↓
	1-3	$p = .001$↑	$p = .010$↑	*ns*
	2-3	$p = .013$↑	$p = .020$↑	$p = .010$↑
Mean length of AS-units	1-2	*ns*	n.s	$p < .001$↓
	1-3	$p < .001$↑	n.s	*ns*
	2-3	$p = .002$↑	n.s	$p < .001$↑
% of error-free AS-units	1-2	$p = .043$↓	*ns*	*ns*
	1-3	$p = .003$↓	*ns*	*ns*
	2-3	*ns*	*ns*	*ns*
Lexical diversity	1-2	*ns*	*ns*	$p = .007$↓
	1-3	$p = .002$↑	*ns*	*ns*
	2-3	$p < .001$↑	$p < .001$↑	$p < .001$↑
Mean length of pauses	1-2	*ns*	*ns*	*ns*
	1-3	$p < .001$↑	$p = .001$↑	$p = .001$↑
	2-3	$p = .009$↑	$p = .001$↑	$p < .001$↑
Number of repairs	1-2	*ns*	*ns*	*ns*
	1-3	*ns*	*ns*	*ns*
	2-3	*ns*	*ns*	*ns*
Mean duration of syllables	1-2	$p = .004$↑	$p = .002$↑	$p = .007$↑
	1-3	$p = .009$↑	*ns*	*ns*
	2-3	*ns*	$p = .001$↓	$p = .001$↓
Mean length of run	1-2	*ns*	*ns*	*ns*
	1-3	*ns*	$p = .010$↑	$p < .001$↑
	2-3	*ns*	$p = .001$↑	$p = .003$↑
Phonation time ratio	1-2	*ns*	*ns*	*ns*
	1-3	*ns*	*ns*	$p = .003$↑
	2-3	$p = .002$↑	*ns*	*ns*

Note. ns = not significant. The alpha level was adjusted as follows: .025 for complexity, .05 for accuracy, .05 for lexical diversity, and .01 for fluency.

AS unit) and oral fluency (longer pauses and longer syllable duration). The teacher-led group increased clauses per AS unit and mean length of run. The teacher and peer group increased mean length of run and phonation time ratio, but they also increased mean pause length.

Research Question 2: The Performances on CALF Between Groups

Research Question 2 asked whether the teacher-led group and the teacher and peer treatment group significantly outperform the comparison group in terms of complexity, accuracy, lexis, and fluency. This research question was answered by conducting a one-way ANOVA to compare the mean differences between groups for Time 1 and ANCOVAs for Time 2 and Time 3. The independent variable was group (3 levels: the comparison group, the teacher-led group, and the teacher and peer group) and the dependent variables were the nine CALF measures (Table 7). The Time 1 measures were the covariate in the ANCOVAs. The one-way ANOVA was conducted for Time 1 in order to assess whether there were significant differences in the groups' performances at the beginning of the study.

A series of ANOVAs and ANCOVAs were used instead of MANOVAs because a preliminary examination of the data indicated that the Pearson correlation among the dependent variables were near zero in some cases (Tables 10, 11, and 12). MANOVA works better with highly negatively correlated dependent variables (Field, 2013; Tabachnick & Fidell, 2007), and it is not effective if correlations among dependent variables are highly positive or uncorrelated. Tabachnick and Fidell (2007) suggested using separate ANOVAs for each dependent variable with a Bonferroni correction as an alternative procedure. Therefore, separate ANOVAs were used to compare the three groups at Time 1 and ANCOVAs were conducted to compare the groups at Time 2 and at Time 3. The covariate was the Time 1 measures.

Before conducing the statistical analyses, the following assumptions for a one-way ANOVA and one-way ANCOVA were checked. First, univariate outliers were checked by converting the raw scores for the speech data to *z*-scores and checking for values > ± 3.29. No univariate outliers were found. Second, normality was checked. One variable, Clauses per AS unit for Time 2 in the teacher and peer group, had skewness and kurtosis statistics > |2.58| (see Table 13). A log transformation was conducted to make it normally distributed (Field, 2013, p. 203). Following the transformation, no *z*-skewness and z-kurtosis values were higher than ± 3.29. I report the results for both the untransformed and transformed values for clauses per AS unit at Time 2 to show if there are any differences in the two data sets. Third, the homogeneity of variance assumption was checked with Levene's test; this assumption was met for all the ANOVAs and ANCOVAs. Lastly, the homogeneity-of-slopes assumption was checked for the one-way ANCOVAs. According to Green and Salkind (2011), a significant interaction between the covariate and the factor suggests that population slopes differ or that the differences on the dependent variable among groups vary as a function of the covariate. If the interaction is significant, ANCOVA should not be used. The homogeneity-of-slopes assumptions were met for all nine ANCOVAs.

First, nine ANOVAs were run to investigate whether there were significant differences in the CALF measures between the three groups at Time 1. The independent variable was group (Three levels: the comparison group, teacher-led group, and teacher and peer group), and the dependent variables were the nine CALF variables; clauses per AS unit, mean length of AS units, error free AS unit, MTLD, mean length of pauses, mean duration of syllable, repair, mean length of run, and phonation time ratio. Using a Bonferroni adjustment for each CALF construct, the alpha level was adjusted as follows: p = .025 for complexity (.05/2), p = .05 for accuracy, p = .05 for lexis, and p = .01 for fluency (.05/5).

Second, nine ANCOVAs were run to investigate whether there were significant differences in the CALF measures between the three groups at Time 2. The independent variable was group (Three levels: the comparison group, teacher-led group, and teacher and peer group), and the dependent variables were clauses per AS unit, mean length of AS units, error free AS unit, MTLD, mean length of pauses, mean duration of syllable, repair, mean length of run, and phonation time ratio. The covariate was the nine measures from Time 1. Using the Bonferroni adjustment for each CALF construct, the alpha level was adjusted as follows: p = .025 for complexity (.05/2), p = .05 for accuracy, p = .05 for lexis, and p = .01 for fluency (.05/5).

Third, nine ANCOVAs were run to investigate whether there were significant differences in the CALF measures between the three groups at Time 3. The independent variable was group (Three levels: the comparison group, teacher-led group, and teacher and peer group), and the dependent variables were clauses per AS unit, mean length of AS units, error free AS unit, MTLD, mean length of pauses, mean duration of syllable, repair, mean length of run, and phonation time ratio. The covariate was the nine measures from Time 1. Using the Bonferroni adjustment for each CALF construct, the alpha level was adjusted as follows: $p = .025$ for complexity (.05/2), $p = .05$ for accuracy, $p = .05$ for lexis, and $p = .01$ for fluency (.05/5).

Clauses per AS unit (complexity). The teacher and peer group produced the most clauses per AS unit ($M = 1.74$, $SD = .28$) at Time 1 as shown in Table 15. The comparison group produced the second most clauses per AS unit ($M = 1.61$, $SD = .31$), and the teacher-led group produced the lowest number of clauses per AS unit ($M = 1.51$, $SD = .20$). A one-way ANOVA was conducted to evaluate if there were significant differences between the three groups. The independent variable was group (Three levels: the comparison, teacher-led, and teacher and peer groups), and the dependent variable was clauses per AS unit at Time 1. The ANOVA was not significant, $F(2, 45) = 2.79$, $p = .07$, partial $\eta^2 = .11$.

The comparison group produced the largest number of clauses per AS unit ($M = 1.78$, $SD = .32$) at Time 2. The teacher and peer group produced fewer clauses per AS unit ($M = 1.51$, $SD = .29$), and the teacher-led group produced the smallest number of clauses per AS unit ($M = 1.48$, $SD = .20$). Clauses per AS unit for the teacher and peer group at Time 2 had skewness or kurtosis statistics > |2.58| (see Table 15). A log transformation was conducted to make the data normally distributed; after the transformation, no z-scores were higher than ± 3.29. A one-way ANCOVA was conducted both with transformed and untransformed data to evaluate if there were significant differences between groups in clauses per AS unit at Time 2. The independent variable was group (Three levels: the comparison, teacher-led, and teacher and peer groups), and the dependent variable was clauses per AS unit. A preliminary analysis evaluating the homogeneity-of-slopes assumption indicated that the relationship between the covariate (clauses per AS unit at Time 1) and the dependent variable (clauses per AS unit at Time 2) did not differ significantly as a function of the independent variable, *F*(2, 42) = .16, *MSE* = .01, *p* = .86, partial η^2 = .007 for untransformed data; *F*(2, 42) = .14, *MSE* = .01, *p* = .87, partial η^2 = .007 for transformed data. For both transformed and untransformed data, the homogeneity of slopes was assumed and the covariate (clauses per AS unit at Time 1) acted similarly across the independent variable (group). The ANCOVA results showed significant differences with the untransformed data, *F*(2, 44) = 4.68, *MSE* = .08, *p* = .01. partial η^2 = .18; and with the transformed data, *F*(2, 44) = 4.74, *MSE* = .08, *p* = .01. partial η^2 = .18. The effect size was medium (Green & Salkind, 2011, p. 213), and the strength of the relationship between group (independent variable) and dependent variable (clauses per AS unit at Time 2) was medium, as assessed by partial η^2, with the group factor accounting for 18% of the variance in the dependent variable. Follow-up tests were conducted to evaluate pairwise differences among group means. Based on the Bonferroni procedure, the adjusted mean for the comparison group differed significantly from the teacher and peer group (*p* = .023).

The comparison group produced the most clauses per AS unit (*M* = 2.13, *SD* = .37) at Time 3. The teacher and peer group produced the second most clauses per AS unit (*M* = 1.81, *SD* = .32), and the teacher-led group produced the fewest clauses per AS unit (*M* = 1.80, *SD* = .33). A one-way ANCOVA was conducted to evaluate if there was a significant difference between the three groups at Time 3. The independent variable was group (Three levels: the comparison, teacher-led, and teacher and peer groups), and the dependent variable was clauses per AS unit. A preliminary analysis evaluating the homogeneity-of-slopes assumption indicated that the relationship between the covariate (clauses per AS unit from Time 1) and the dependent variable (clauses per AS unit at Time 3) did not differ significantly as a function of the independent variable, *F*(2, 42) = .64, *MSE* = .12, *p* = .53, partial η^2 = .03. The ANCOVA was significant, *F*(2, 42) = 4.55, *MSE* = .11, *p* = .016. partial η^2 = .17. Based on the Bonferroni adjustment, the comparison group differed significantly from the teacher and peer group (*p* = .021).

Mean length of AS units (complexity). The teacher and peer group produced the longest AS unit (*M* = 10.86, *SD* = 1.80) at Time 1. The comparison group had the second longest AS unit (*M* = 10.62, *SD* = 2.01), and the teacher-led group produced the shortest AS unit (*M* = 10.51, *SD* = 1.25). A one-way ANOVA was conducted to evaluate if there was a significant difference between groups. The independent variable was group (Three levels: the comparison, teacher-led, and teacher and peer groups), and the dependent variable was mean length of AS units at Time 1. The ANOVA was not significant, *F*(2, 45) = 1.00, *p* = .37, partial η^2 = .04.

The comparison group produced the longest AS unit (*M* = 10.81, *SD* = 1.99) at Time 2. The teacher-led group produced the second longest AS unit (*M* = 9.57, *SD* = 2.20), and the teacher and peer group produced the shortest AS unit (*M* = 9.38, *SD* = 1.27). A one-way ANCOVA was conducted. The independent variable was group (Three levels: the comparison, teacher-led, and teacher and peer groups), and the dependent variable was mean length of AS units at Time 2. A preliminary analysis evaluating the homogeneity-of-slopes assumption indicated that the relationship between the covariate (mean length of AS units at Time 1) and the dependent variable (mean length of AS units at Time 2) did not differ significantly as a function of the independent variable, *F*(2, 42) = .80, *MSE* = 3.10, *p* = .46, partial η^2 = .04. The ANCOVA was not significant, *F*(2, 44) = 2.81, *MSE* = 3.17, *p* = .07. partial η^2 = .11.

At Time 3, the comparison group produced the longest AS unit (*M* = 13.63, *SD* = 1.87), the teacher-led group produced the second longest AS unit (*M* = 11.43, *SD* = 2.00), and the teacher and peer group produced the shortest AS unit (*M* = 11.15, *SD* = 1.88). A one-way ANCOVA was conducted. The independent variable was group (Three levels: the comparison, teacher-led, and teacher and peer groups), and the dependent variable was mean length of AS units at Time 3. A preliminary analysis evaluating the homogeneity-of-slopes assumption indicated that the relationship between the covariate (mean length of AS unit from Time 1) and the dependent variable (mean length of AS units at Time 3) did not differ significantly as a function of the independent variable, *F*(2, 42) = .39, *MSE* = 3.52, *p* = .68, partial η^2 = .02. The ANCOVA was significant, *F*(2, 44) = 7.91, *MSE* = 3.42, *p* = .001, partial η^2 = .27. The effect size was large (Green & Salkind, 2011, p. 213). The strength of the relationship between group (independent variable) and the dependent variable (mean length of AS units at Time 3) was strong, as assessed by a partial η^2, with the group factor accounting for 27% of the variance in the dependent variable.

The comparison group had the largest adjusted mean (*M* = 13.60), the teacher-led group had a smaller adjusted mean (*M* = 11.60), and the teacher and peer group had the smallest adjusted mean (*M* = 11.05). Follow-up tests were conducted to evaluate pairwise differences among the group means. Based on the Bonferroni procedure, the adjusted means for the comparison group differed significantly from the teacher and peer group (*p*

< .001). The adjusted mean for the comparison group differed significantly from the teacher-led group (*p* = .008). However, the adjusted means for the two experimental groups did not differ significantly (*p* = .40).

Accuracy. The comparison group had the highest mean for accuracy at Time 1 (*M* = .71, *SD* = .16), showing that 71% of the AS units were correct forms. The teacher and peer group had the second highest mean for error-free AS units (*M* = .66, *SD* = .17). The teacher-led group had the least mean for error-free AS units (*M* = .60, *SD* = .12). A one-way ANOVA was conducted to evaluate if there was a significant difference between the three groups. The independent variable was group (Three levels: the comparison, teacher-led, and teacher and peer groups), and the dependent variable was error-free AS units at Time 1. The ANOVA was not significant, *F*(2, 45) = 1.78, *p* = .18, partial η^2 = .07.

The teacher-led group had the highest mean for accuracy at Time 2 (*M* = .64, *SD* = .20), showing that 64% of the AS units were correct forms. The teacher and peer group had the second highest mean for error-free AS units (*M* = .62, *SD* = .24), and the comparison group had the lowest mean for error-free AS units (*M* = .60, *SD* = .19). A one-way ANCOVA was conducted. The independent variable was group (Three levels: comparison, teacher-led, and teacher and peer groups), and the dependent variable was error-free AS units at Time 2. A preliminary analysis evaluating the homogeneity-of-slopes assumption indicated that the relationship between the covariate (error-free AS units at Time 1) and the dependent variable (error-free AS units at Time 2) did not differ significantly as a function of the independent variable, *F*(2, 42) = 1.33, *MSE* = .05, *p* =

.28, partial η^2 = .06. The ANCOVA was not significant, *F*(2, 44) = .34, *MSE* = .05, *p* = .71. partial η^2 = .02.

The teacher-led group had the highest mean for accuracy at Time 3 (*M* = .73, *SD* = .16), showing that 73% of AS units were correct forms. The teacher and peer group had the second highest mean for error-free AS units (*M* = .63, *SD* = .19), and the comparison group had the lowest mean for error-free AS units (*M* = .59, *SD* = .17). A one-way ANCOVA was conducted. The independent variable was group (Three levels: the comparison, teacher-led, teacher and peer groups), and the dependent variable was error-free AS units at Time 3. A preliminary analysis evaluating the homogeneity-of-slopes assumption indicated that the relationship between the covariate (error-free AS units from Time 1) and the dependent variable (error-free AS units at Time 3) did not differ significantly as a function of the independent variable, *F*(2, 42) = 1.67, *MSE* = .03, *p* = .20, partial η^2 = .07. The ANCOVA was not significant, *F*(2, 44) = 3.41, *MSE* = .03, *p* = .04. partial η^2 = .13.

Lexis. The teacher-led group had the highest mean score at Time 1 (*M* = 37.62, *SD* = 9.16), the teacher and peer group had a slightly lower MTLD (*M* = 37.38, *SD* = 10.56), and the comparison group had the lowest MTLD (*M* = 34.47, *SD* = 8.32). A one-way ANOVA was run to evaluate if there was a significant difference between groups. The independent variable was group (Three levels: the comparison, teacher-led, and teacher and peer groups), and the dependent variable was MTLD at Time 1. The ANOVA was not significant, *F*(2, 45) = .47, *p* = .63. partial η^2 = .02.

The teacher-led group had the highest mean score at Time 2 (*M* = 33.60, *SD* = 7.40), the teacher and peer group had a lower mean (*M* = 30.04, *SD* = 7.36), and the comparison group had the lowest mean for MTLD (*M* =30.03, *SD* = 4.75). A one-way ANCOVA was conducted. The independent variable was group (Three levels: the comparison, teacher-led, and teacher and peer groups), and the dependent variable was MTLD at Time 2. A preliminary analysis evaluating the homogeneity-of-slopes assumption indicated that the relationship between the covariate (MTLD from Time 1) and the dependent variable (MTLD at Time 2) did not differ significantly as a function of the independent variable, *F*(2, 42) = .18, *MSE* = 47.06, *p* = .84, partial η^2 = .008. The ANCOVA was not significant, *F*(2, 44) = 1.26, *MSE* = 45.30, *p* = .29, partial η^2 = .05.

The comparison group had the highest mean score (*M* = 46.26, *SD* = 10.16) at Time 3. The teacher-led group had slightly lower MTLD (*M* = 44.06, *SD* = 9.61), and the comparison group had the lowest mean for MTLD (*M* = 41.82, *SD* = 11.27). A one-way ANCOVA was conducted. The independent variable was group (Three levels: the comparison, teacher-led, and teacher and peer groups), and the dependent variable was MTLD at Time 3. A preliminary analysis evaluating the homogeneity-of-slopes assumption indicated that the relationship between the covariate (MTLD from Time 1) and the dependent variable (MTLD at Time 3) did not differ significantly as a function of the independent variable, *F*(2, 42) = .34, *MSE* = 110.48, *p* = .71, partial η^2 = .02. The ANCOVA was not significant, *F*(2, 44) = .99, *MSE* = 107.17, *p* = .38, partial η^2 = .04.

Mean length of pauses (fluency). The teacher and peer group had the shortest mean pauses (*M* = .59, *SD* = .18) at Time 1, the teacher-led group had the second shortest pauses (*M* = .70, *SD* = .18), and the comparison group had the longest pauses (*M* = .75, *SD* = .21). A one-way ANOVA was conducted to evaluate if there was a significant difference between groups. The independent variable was group (Three levels: the comparison, teacher-led, and teacher and peer groups), and the dependent variable was mean length of pauses at Time 1. The ANOVA was not significant, *F*(2, 45) = 4.09, *p* = .02. partial η^2 = .15.

The teacher and peer group had the shortest mean pauses at Time 2 (*M* = 62.41, *SD* = .17), the teacher-led group had had the second shortest pauses (*M* = .72, *SD* = .20), and the comparison group had the longest pauses (*M* = .86, *SD* = .13). A one-way ANCOVA was conducted. The independent variable was group (Three levels: the comparison, teacher-led, and teacher and peer groups), and the dependent variable was mean length of pauses at Time 2. A preliminary analysis evaluating the homogeneity-of-slopes assumption indicated that the relationship between the covariate (mean length of pauses from Time 1) and the dependent variable (mean length of pauses at Time 2) did not differ significantly as a function of the independent variable, *F*(2, 42) = .58, *MSE* = .02, *p* = .56, partial η^2 = .03. The ANCOVA was not significant, *F*(2, 44) = 4.00, *MSE* = .02, *p* = .03, partial η^2 = . 15.

The comparison group had the shortest mean pauses at Time 3 (*M* = 1.05, *SD* = .16), the teacher-led group had the second shortest pauses (*M* = 1.31, *SD* = .38), and the teacher and peer group had the longest pauses (*M* = 1.58, *SD* = .43). A one-way ANCOVA was conducted. The independent variable was group (Three levels: the comparison, teacher-led, and teacher and peer groups), and the dependent variable was mean length of pauses at Time 3. A preliminary analysis evaluating the homogeneity-of-slopes assumption indicated that the relationship between the covariate (mean length of pauses from Time 1) and the dependent variable (mean length of pauses at Time 3) did not differ significantly as a function of the independent variable, $F(2, 42) = 2.24$, $MSE = .26$, $p = .12$, partial $\eta^2 = .10$. The ANCOVA was significant, $F(2, 44) = 5.64$, $MSE = .12$, $p = .007$, partial $\eta^2 = .20$. Based on the Bonferroni procedure, the comparison group differed significantly from the teacher and peer group ($p = .005$), meaning that the comparison groups had much shorter pause length.

Repairs (fluency). The teacher and peer group had the lowest mean score ($M = 7.71$, $SD = 5.03$) at Time 1, suggesting that the group did not repeat or self-correct many times. The teacher-led group had the second lowest repair scores ($M = 10.50$, $SD = 5.88$) followed by the comparison group ($M = 12.15$, $SD = 9.07$). A one-way ANOVA was conducted to evaluate if there was a significant difference between groups. The independent variable was group (Three levels: the comparison, teacher-led, and teacher and peer groups), and the dependent variable was repairs at Time 1. The ANOVA was not significant, $F(2, 45) = 1.97$, $p = .15$, partial $\eta^2 = .08$.

The teacher and peer group had the lowest mean score ($M = 7.00$, $SD = 4.52$) at Time 2, indicating that the group did not repeat or self-correct frequently. The teacher-led group had the second lowest repair scores ($M = 12.07$, $SD = 7.93$), and the comparison group had the most frequent repairs ($M = 12.31$, $SD = 4.70$). A one-way ANCOVA was conducted. The independent variable was group (Three levels: the comparison, teacher-led, and teacher and peer groups), and the dependent variable was repairs at Time 2. A preliminary analysis evaluating the homogeneity-of-slopes assumption indicated that the relationship between the covariate (repairs from Time 1) and the dependent variable (repairs at Time 2) did not differ significantly as a function of the independent variable, *F*(2, 42) = 3.60, *MSE* = 21.55, *p* = .04, partial η^2 = .15. The ANCOVA was not significant, *F*(2, 44) = 2.85, *MSE* = 24.09, *p* = .07, partial η^2 = .12.

The teacher and peer group had the lowest mean score at Time 3 (*M* = 8.05, *SD* = 5.02), suggesting that the group did not repair frequently. The teacher-led group produced the second lowest repair scores (*M* = 10.64, *SD* = 5.72), and the comparison group produced the most frequent repairs (*M* = 11.69, *SD* = 2.92). A one-way ANCOVA was conducted. The independent variable was group (Three levels: the comparison, teacher-led, and teacher and peer groups), and the dependent variable was repairs at Time 3. A preliminary analysis evaluating the homogeneity-of-slopes assumption indicated that the relationship between the covariate (Repairs from Time 1) and the dependent variable (Repairs at Time 3) did not differ significantly as a function of the independent variable, *F*(2, 42) = 1.79, *MSE* = 19.15, *p* = .18, partial η^2 = .08. The ANCOVA was not significant, *F*(2, 44) = 1.24, *MSE* = 19.84, *p* = .30, partial η^2 = .05.

Mean duration of syllable (fluency). The comparison group had the lowest mean score (*M* =.28, *SD* = .03) at Time 1; thus, they spent shortest time to pronounce syllables. The teacher-led group had the second lowest mean score (*M* = .29, *SD* = .04), and the comparison group had the highest score (*M* = .30, *SD* = .05). A one-way ANOVA was conducted to evaluate if there were significant differences between the three groups. The independent variable was group (Three levels: the comparison, teacher-led, and teacher and peer groups), and the dependent variable was mean duration of syllable at Time 1. The ANOVA was not significant, *F*(2, 45) = 1.53, *p* = .23, partial η^2 = .06.

The comparison group had the lowest mean score at Time 2 (*M* = .317, *SD* = .04), meaning that the group spent the shortest time to pronounce a syllable among the three groups. The teacher-led group had a similar mean score (*M* = .318, *SD* = .03), and the teacher and peer group had the highest score (*M* = .334, *SD* = .04). A one-way ANCOVA was conducted. The independent variable was group (Three levels: the comparison, teacher-led, and teacher and peer groups), and the dependent variable was mean duration of syllable at Time 2. A preliminary analysis evaluating the homogeneity-of-slopes assumption indicated that the relationship between the covariate (mean duration of syllable from Time 1) and the dependent variable (mean duration of syllable at Time 2) did not differ significantly as a function of the independent variable, *F*(2, 42) = .29, *MSE* = .001, *p* = .75, partial η^2 = .01. The ANCOVA was not significant, *F*(2, 42) = .49, *MSE* = .001, *p* = .02., partial η^2 = .02.

The teacher-led group had the lowest mean score at Time 3 (*M* =.287, *SD* = .033), meaning that the group spent the shortest time to pronounce a syllable among the three groups. The comparison group had the second lowest mean score (*M* =.304, *SD* = .031), and the teacher and peer group had the highest score (*M* = .306, *SD* = .031). A one-way ANCOVA was conducted. The independent variable was group (Three levels: the comparison, teacher-led, and teacher and peer groups), and the dependent variable was mean duration of syllable at Time 3. A preliminary analysis evaluating the homogeneity-of-slopes assumption indicated that the relationship between the covariate (mean duration of syllable from Time 1) and the dependent variable (mean duration of syllable at Time

3) did not differ significantly as a function of the independent variable, *F*(2, 42) = 2.40, *MSE* = .001, *p* = .10, partial η^2 = .10. The ANCOVA was not significant, *F*(2, 44) = 1.63, *MSE* = .001, *p* = .21, partial η^2 = .07.

Mean length of run (fluency). The comparison group had the longest mean length of run among three groups (*M* = 4.88, *SD* = .75) at Time 1. The teacher-led group had the second longest mean length of run (*M* = 4.58, *SD* = .72), followed by the teacher and peer group (*M* = 3.95, *SD* = .76). A one-way ANOVA was conducted to evaluate if there was a significant difference between the three groups. The independent variable was group (Three levels: the comparison, teacher-led, and teacher and peer groups), and the dependent variable was mean length of run at Time 1. The ANOVA was significant, *F*(2, 45) = 6.81, *p* = .003, partial η^2 = .23. The strength of the relationship between groups and the mean length of run as assessed by η^2 was strong, with the group factor accounting for 23% of the variance in the dependent variable. Follow-up tests were conducted with a Bonferroni test, as it is a conservative test that controls for Type I error. There was a significant difference in the means between the comparison group and the teacher and peer group (*p* = .003) but no significant difference between the teacher-led group and the teacher and peer group. The comparison group showed a longer mean length of run in comparison to the teacher and peer group.

The teacher and peer group had the longest mean length of run (*M* = 4.50, *SD* = 1.18) at Time 2, the comparison group had the second longest mean length of run (*M* = 4.34, *SD* = .78), and the teacher-led group had the shortest mean length of run (*M* = 4.33, *SD* = .84). A one-way ANCOVA was conducted. The independent variable was group

(Three levels: the comparison, teacher-led, and teacher and peer groups), and the dependent variable was mean length of run at Time 2. A preliminary analysis evaluating the homogeneity-of-slopes assumption indicated that the relationship between the covariate (mean length of run from Time 1) and the dependent variable (mean length of run at Time 2) did not differ significantly as a function of the independent variable, *F*(2, 42) = .62, *MSE* = .74, *p* = .54, partial η^2 = .03. The ANCOVA was not significant, *F*(2, 44) = 3.24, *MSE* = .73, *p* = .05, partial η^2 = .13.

The teacher-led group had the longest mean length of run at Time 3 (*M* = 5.31, *SD* = .86), the teacher and peer group had the second longest mean length of run (*M* = 5.14, *SD* = .93), and the comparison group had the shortest mean length of run (*M* = 5.05, *SD* = .80). A one-way ANCOVA was conducted. The independent variable was group (Three levels: the comparison, teacher-led, and teacher and peer groups), and the dependent variable was mean length of run at Time 3. A preliminary analysis evaluating the homogeneity-of-slopes assumption indicated that the relationship between the covariate (mean length of run from Time 1) and the dependent variable (mean length of run at Time 3) did not differ significantly as a function of the independent variable, *F*(2, 42) = .10, *MSE* = .70, *p* = .91, partial η^2 = .005. The ANCOVA was not significant, *F*(2, 44) = 1.31, *MSE* = .67, *p* = .28, partial η^2 = .06.

Phonation time ratio (fluency). The comparison group had the highest mean score among groups at Time 1 (*M* = 53.97, *SD* = 8.50), indicating that the group spent approximately 53.98% of time speaking. The teacher-led group had the second highest phonation time ratio (*M* = 52.31, *SD* = 9.65) followed by the teacher and peer group (*M* =

45.50, *SD* = 6.11). The ANOVA was significant, *F*(2, 45) = 5.60, *p* = .007, partial η^2 = .19; the Bonferroni post hoc test indicated that the comparison group differed significantly from the teacher and peer group at Time 1.

The comparison group had the highest mean score among the groups at Time 2 (*M* = 55.19, *SD* = 5.90); the group spoke 55.19% of the time. The teacher-led group had the second highest phonation time ratio (*M* = 49.68, *SD* = 9.59) followed by the teacher and peer group (*M* = 48.48, *SD* = 10.44). A one-way ANCOVA was conducted. The independent variable was group (Three levels: the comparison, teacher-led, and teacher and peer groups), and the dependent variable was phonation time ratio at Time 2. A preliminary analysis evaluating the homogeneity-of-slopes assumption indicated that the relationship between the covariate (phonation time ratio from Time 1) and the dependent variable (phonation time ratio at Time 2) did not differ significantly as a function of the independent variable, *F*(2, 42) = 1.08, *MSE* = 57.66, *p* = .35, partial η^2 = .05. The ANCOVA was not significant, *F*(2, 44) = 1.27, *MSE* = 57.88, *p* = .29, partial η^2 = .06.

The comparison group had the highest mean score among groups at Time 3 (*M* = 60.83, *SD* = 4.33); the group spoke 60.83% of the time. The teacher-led group had the second highest phonation time ratio (*M* = 55.61, *SD* = 7.85) followed by the teacher and peer group (*M* = 51.85, *SD* = 8.88). A one-way ANCOVA was conducted. The independent variable was group (Three levels: the comparison, teacher-led, and teacher and peer groups), and the dependent variable was phonation time ratio at Time 3. A preliminary analysis evaluating the homogeneity-of-slopes assumption indicated that the relationship between the covariate (phonation time ratio from Time 1) and the dependent variable (phonation time ratio at Time 3) did not differ significantly as a function of the independent variable, $F(2, 42) = .76$, $MSE = 49.84$, $p = .47$, partial $\eta^2 = .04$. The ANCOVA was not significant, $F(2, 44) = 2.38$, $MSE = 49.29$, $p = .10$, partial $\eta^2 = .19$.

Table 45: *Summary of Between-Group Differences*

	Time 1 ANOVA	Time 2 ANCOVA	Time 3 ANCOVA
Complexity			
Clauses per AS-unit	*ns*	*Comparison group > T & P group	* Comparison group > Teacher and peer group
Mean length of AS units	*ns*	*ns*	* Comparison group > Teacher and peer group * Comparison group > Teacher-led group
Accuracy			
% of error-free AS-units	*ns*	*ns*	*ns*
Lexical diversity	*ns*	*ns*	*ns*
Fluency			
Mean length of pauses	*ns*	*ns*	* Comparison group < Teacher and peer group
Number of repairs	*ns*	*ns*	*ns*
Mean duration of syllables	*ns*	*ns*	*ns*
Mean length of run	* Comparison group > Teacher and peer group	*ns*	*ns*
Phonation time ratio	* Comparison group > Teacher and peer group	*ns*	*ns*

Note. *Statistically significant.

In sum, the statistical analyses indicated that the comparison group was more fluent than the teacher and peer group as indicated by mean length of run and phonation time ratio at Time 1. ANCOVAs were conducted with the Time 1 measures as the covariate to compare group differences at Time 2 and Time 3. The comparison group produced more clauses per AS-unit, longer AS-unit, and shorter pauses compared to the teacher and peer group at Time 3. Table 45 shows the summary of the findings for research question 2.

Research Question 3: The Frequency and Types of Target Form

Research question 3 asked to what extent the participants who received form-focus pedagogic intervention used the target formulaic language in terms of frequency and variety across the 13 weeks. The frequency and the variety of the target formulaic language such as *in my opinion*, *it is mainly because…* or *for example…* was analyzed. An ANOVA was not conducted because the descriptive analysis gives a clear picture of the participants' use of the formulaic language given that the data are in the form of raw frequencies. The average usage of the target formulaic language per person was calculated by dividing the total number of occurrences by the number of participants in the group. For example, if the comparison group used *in my opinion* 20 times, the 20 occurrences was divided by 13 (i.e., there were 13 participants in that group) to arrive at an average number. That number was then used to make comparisons with the other groups' use of the same target formulaic language.

In addition to counting the frequency and variety of the target formulaic language, I showed transcriptions supporting the statistical evidence. The repeated listening and transcribing procedures revealed several distinctive characteristics in the speakers' monologues and helped explain how the participants organized their monologues. Excerpts were selected from nine speakers because their use of the target formulaic language changed a great deal from Time 1 to Time 3. These data illustrate how they used the target formulaic language more effectively or more inappropriately over the academic semester.

Opinion Phrases

Table 46 shows the descriptive statistics of the target formulaic language that the participants used to express opinions at Time 1, Time 2, and Time 3. The target phrase *in my opinion* was used more frequently than the other opinion phrases such as *personally speaking, I think*, or *I'm not sure, but I think* throughout the semester. None of the participants used the target formulaic language to express opinions at Time 1, notwithstanding the fact that they had already learned the target formulaic language for opinions in the lesson before the Time 1 recording.

The teacher and peer group used the target phrase *In my opinion* more frequently than other two groups at Time 2 (the teacher and peer group: 17 users, 20 times, *M* = .95; the comparison group: 3 users, 3 times, *M* = .23; the teacher-led group: 5 users, 5 times, *M* = .36). The differences in the use of the formulaic language between the experimental groups and the comparison group became more distinct at Time 3. Both the teacher-led and the teacher and peer group used *in my opinion* more frequently than the comparison group at Time 3. For instance, the comparison group used *in my opinion* only 2 times in total (*M* = .15) and they never used other formulaic language such as *personally speaking, I think* and *I'm not sure, but I think.* On the other hand, 11 students from the teacher-led group used *in my opinion* 18 times (*M* = 1.29) and *I'm not sure, but I think* once (*M* = .007). The teacher and peer group used a wider variety of target phrases than

Table 46. *Descriptive Statistics of Types of Formulaic Language Used for Opinion Function*

		Opinion								
		In my opinion			Personally speaking, I think			I am not sure, but I think		
Group		Time 1	Time 2	Time 3	Time 1	Time 2	Time 3	Time 1	Time 2	Time 3
Comparison	Users	3	3	2	0	1	0	0	2	0
(*n* = 12)	Counts	3	3	2	0	1	0	0	2	0
	M	0.23	0.23	0.15	0.00	0.08	0.00	0.00	0.15	0.00
Teacher-led	Users	2	5	11	0	2	0	0	0	1
(*n* = 13)	Counts	2	5	18	0	2	0	0	0	1
	M	0.14	0.36	1.29	0.00	0.14	0.00	0.00	0.00	0.07
Teacher and peer	Users	4	17	19	0	0	4	3	4	4
(*n* = 21)	Counts	4	20	20	0	0	5	3	4	4
	M	0.19	0.95	0.95	0.00	0.00	0.24	0.14	0.19	0.19

Note. Users = the number of users who used the formulaic language. Counts = frequency of the formulaic language occurrences. *M* = total count/number of participants in the group.

the other two groups. The teacher and peer group used *in my opinion* 20 times (19 users, *M* = .95) as well as *personally speaking, I think* 5 times (4 users, *M* = .24) and *I'm not sure, but I think* 4 times (4 users, *M* = .19).

The participants in the comparison group did not frequently use the target formulaic language for opinions throughout the semester, in part because they repeated *I think* to state their opinions instead of using the target opinion phrases. One example from the comparison group at Time 1 is as follows:

Excerpt (1) Student 1: Comparison Group at Time 1

1. *I think* :: doing club activity is a good idea for students :: because we can many experiences in club activities
2. I will join movie club in this university :: because movie circle will be make good relations for me
3. I didn't join a club in my high school :: but I wanted to join a club

The speaker above used *I think* (line 1) to state her opinion concerning a club activity at Time 1. The following transcript shows a portion of the same participant's monologue at Time 3.

Excerpt (2) Student 1: Comparison Group at Time 3

1. *I think* :: study abroad is a good way for university students :: because if we go to study abroad :: we have to speak in English every time
2. the opportunity :: which speak Japanese :: is decreasing
3. so we can communicate with others in English more flexible
4. so study abroad is a good idea for university students

The phrase *I think* is an appropriate way to state an opinion. Although the comparison group learned the target formulaic language outside of the treatment phase and they were able to use the phrases when they engaged in group discussions, many of them did not apply what they learned outside of the treatment phase in their 3/2/1 monologue.

The participants in the teacher and peer group increased the frequency and the variety of target formulaic language. One participant in the teacher and peer group did not explicitly state her opinion at Time 1, as shown in excerpt (3).

Excerpt (3) Student 2: Teacher and Peer Group at Time 1

1. I want :: to join ski and water club in R university
2. but this circle is so spend so much money
3. I give up join this circle and
4. I belong to volleyball circle now :: because I played volleyball :: when I was a high school and junior high school student
5. I want :: to play something sports :: and I’m good at volleyball
6. but I want :: to do something new thing :: but I can't find good circle

The participant’s monologue is understandable to some extent, but she did not state her opinion that doing a club activity is a good idea. In contrast, she clearly stated her opinion in her monologue at Time 3 as shown in excerpt (4).

Excerpt (4) Student 2: Teacher and Peer Group at Time 3

1. *In my opinion* I think :: learning English is important for us
2. *It’s mainly because* :: now foreign people are increasing in Japan :: and we should speak with them or communicate with them :: and we should live with them

The participants in the teacher and peer group used a wider variety of the target formulaic language compared to the other two groups. Another female participant in the teacher and peer group stated her opinion as follows at Time 1. She repeatedly said *I think* without using the target formulaic language as shown in excerpt (5).

Excerpt (5) Student 3: Teacher and peer group at Time 1

1. *I think* :: doing club activities is a good idea for students
2. So it's important for students :: to make important mates
3. So *I think* :: they are always be good terms
4. So there's sometimes makes conflict with them
5. But they can understand each other more than other club mates
6. So they will be good partner in the future
7. I decided my circle in R activity :: because I'm interested in volunteer works :: before I become a university student
8. So I liked :: to make someone smile so
9. I want :: to talk many people and
10. I enjoy to play with many people

The same participant used a variety of phrases such as *in my opinion*, *I'm not sure, but I think*, and *personally speaking, I think* at Time 3. Excerpt (6) shows how much student 3 increased her use of the target formulaic language. She used many opinion phrases, but she did not elaborate on her ideas with detailed reasons and examples. Nevertheless, her opinions are clearly expressed at Time 3. In contrast, no one in the comparison group used the target formulaic language for expressing opinions even at Time 3.

Excerpt (6) Student 3: Teacher and peer group at Time 3

1. <u>*In my opinion*</u> learning English is important for me
2. <u>*I'm not sure :: but I think*</u> :: actually English is used all over the world
3. So *if* I go abroad :: I can speak only English :: but I can communicate with these people in there
4. And <u>*I'm not sure :: but I think*</u> making foreign friends is good way :: to improve my English skill
5. <u>*Personally speaking I think*</u> :: listen in fluently English is good way :: to improve my English skill
6. So I have a half friend :: and she can speak English very well
7. So when I meet her :: she speaks very in fluently English :: and I listen her English
8. My ear can listen English and

Reason Phrases

Table 47 shows the descriptive statistics for the target formulaic language that the participants used to give reasons at Time 1, Time 2, and Time 3. The participants repeatedly used *because* to give reasons instead of the target reason phrases at Time 1. No group used the target formulaic language such as *it's mainly because…*, *one reason is…*, and *another reason is...* at Time 1.

The comparison group continued to use *because…* to give reasons throughout the academic semester (Time 1 = 16 counts, Time 2 = 19 counts, Time 3 = 21 counts). The comparison group's use of other reason phrases such as *one reason is…* (3 counts) and *another reason is...* (6 counts) was not as frequent as the other groups at Time 2. The comparison group used only *because* or *it's mainly because...* at Time 3; they did not use other forms to give reasons.

The teacher-led group also continued to use *because...* and did not use the target formulaic language at Time 1 and Time 2. On the contrary, nine participants in the teacher-led group increased their use of *it's mainly because…* (18 counts, *M* = 1.29) at Time 3 but they did not use other target formulaic language such as *one reason is…* or *another reason is…* at Time 3.

The teacher and peer group used a wider variety of target phrases than the other two groups. They started increasing their use of the target formulaic language to give reasons from Time 2. Six speakers used target phrases such as *one reason is...* (6 counts, *M* = .29) and eight speakers used *another reason is...* (10 counts, *M* = .48) at Time 2. Seven speakers used the target phrases *one reason is…* (7 counts, *M* = .33), and *another reason is...* (7 counts, *M* = .33) at Time 3. The following excerpt is from a male participant in the teacher and peer group at Time 1. This participant gave a reason when he said "*because* club activity can take that a lot of experiences or the skill up or skill up and hobbies and make many friends" at Time 1 (line 1). In this line, several elements such as *experiences, skill, hobby and friendship* constitute reasons. Because he told all the reasons at once, even if he had given follow-up examples, the way he gave reasons prevented him from producing coherently organized discourse.

Excerpt (7) Student 4: Teacher and Peer Group at Time 1

1. I think :: club activity is a good idea for students :: <u>*because*</u> club activity can take that a lot of experiences or the skill up or skill up and hobbies and make many friends
2. Make many friends is very important in school lives
3. So club activity is a good chance :: to make friends and
4. I join the club climbing mountains and broadcast club
5. Climbing mountains clubs is my hobby
6. I can skill up the climbing mountain experiences or skills
7. And so broadcast club is not my hobby
8. But I can skill up using camera or using electrics or voices so

Student 4 increased his usage of the target formulaic language at Time 3 by using the target formulaic language *it's mainly because...* (lines 2, 7). Student 4's main reason why learning English is important was because he would be able to read foreign literature in the original language (line 2). His approach to giving reasons is more coherent here than in the previous speech (excerpt 7).

Table 47. *Descriptive Statistics of Types of Formulaic Language Used for Reason Function*

		Reason											
		because			It's mainly because			One reason is			Another reason is		
Group		Time 1	Time 2	Time 3	Time 1	Time 2	Time 3	Time 1	Time 2	Time 3	Time 1	Time 2	Time 3
Comparison	Users	10	9	13	0	1	5	0	3	0	0	3	0
n = 12	Count	16	19	21	0	1	10	0	3	0	0	6	0
	M	1.23	1.46	1.62	0.00	0.08	0.77	0.00	0.23	0.00	0.00	0.46	0.00
Teacher-led	Users	9	10	1	0	3	9	0	0	0	0	1	0
n = 13	Count	10	14	7	0	3	18	0	0	0	0	1	0
	M	0.71	1.00	0.50	0.00	0.21	1.29	0.00	0	0.00	0.00	0.07	0.00
Teacher and peer	Users	17	7	5	0	10	15	0	6	7	0	8	7
n = 21	Count	25	10	5	0	12	19	0	6	7	0	10	7
	M	1.19	0.48	0.24	0.00	0.57	0.90	0.00	0.29	0.33	0.00	0.48	0.33

Note. Users = the number of users who used the formulaic language. Counts = frequency of the formulaic language occurrences. *M* = total count/number of participants in the group.

Excerpt (8) Student 4: Teacher and Peer Group at Time 3

1. Learning English is important for me
2. *It's mainly because* :: I want to read the books that in English such as Australian history or American history
3. So it's may be translated in Japanese :: but I want :: to read that native language
4. Also I want :: to go abroad easily
5. So learning English is important
6. And studying abroad is a good idea for university students *in my opinion*
7. *It's mainly because* :: foreign students speak English natively

Another example of how the participants in the teacher and peer group developed their use of the target formulaic language is shown in excerpt (9), which is from one female participant at Time 1. She expressed her opinion (line 1), but she did not explicitly use the reason phrases.

Excerpt (9) Student 5: Teacher and Peer Group at Time 1

1. I agree :: with doing club activities is a good idea for students
2. I have join a club activity before
3. I learned from my experiences
4. First I think :: that I can make friends
5. And I know about my friends' university and classes
6. And we help each other

The same participant was able to give clearer reasons by using *one reason is…* (Excerpt 10, line 2) and *another reason is...* (line 4) at Time 3. Although each reason is short—"*English is very useful language*" (line 2), and "*I can learn other culture*" (line 4)—her use of the target formulaic language made her reasons easier to understand.

Excerpt (10) Student 5: Teacher and Peer Group at Time 3

1. In my opinion learning English is important for me
2. *One reason is* :: English is very useful language
3. So I can communicate with many foreign people and
4. *Another reason is* :: I can learn other culture
5. I know other culture's good points and bad points :: and I can know Japanese culture's good points and bad points

Example Phrases

Table 48 shows the descriptive statistics for the target formulaic language the participants used to give examples at Time 1, Time 2, and Time 3. First, both the teacher- led and the teacher and peer groups used example phrases much more frequently than the comparison group. Eight speakers in the teacher-led group used *for example* (10 counts, *M* = .71), while 15 speakers in the teacher and peer group did so (23 counts, *M* = 1.10) at Time 3. In contrast, two participants in the comparison group only gave examples twice (*M* = .15) at Time 3. This result is surprising given that *for example* is a phrase that Japanese students learn in junior high school. This finding suggests that the participants in the experimental group learned to use the target formulaic language for giving examples as a result of the pedagogical intervention. Second, no participants used a variety of formulaic language for giving examples. The most frequently used formulaic language was *for example*; other target phrases such as *one example is…* or *another example is…* were not used as frequently.

Table 48. *Descriptive Statistics of Types of Formulaic Language Used for Example Function*

					Example					
			For example			One example is…			Another example is...	
		Time 1	Time 2	Time 3	Time 1	Time 2	Time 3	Time 1	Time 2	Time 3
Group										
Comparison	Users	0	6	2	0	0	0	0	0	0
n = 12	Counts	0	7.00	2.00	0	0	0	0	0	0
	M	0.00	0.54	0.15	0.00	0.00	0.00	0.00	0.00	0.00
Teacher-led	Users	2	5	8	0	0	1	0	0	1
n = 13	Counts	2	5	10	0	0	1	0	0	1
	M	0.14	0.36	0.71	0.00	0.00	0.07	0	0	0.07
Teacher and peer	Users	4	10	15	0	1	0	0	1	0
n = 21	Counts	4	13	23	0	2	0	0	1	0
	M	0.19	0.62	1.10	0.00	0.10	0.00	0.00	0.05	0.00

Note. Users = the number of users who used the formulaic language. Counts = frequency of the formulaic language occurrences. *M* = total count/number of participants in the group.

Excerpts 11 and 12 show a speaker in the comparison group who did not give examples at Time 2 and Time 3. One female participant in the comparison group never used the target example phrases at Times 1, 2, and 3.

Excerpt (11) Student 6: Comparison Group at Time 2

1. I think :: eat in is more time than eat out
2. So when I eat in :: I eat dinner with my family
3. And when I eat out :: I eat dinner with my friend or alone
4. But I think :: eat in is better than eat out

Student 6 could have elaborated on how she enjoyed eating with her family and friends by giving examples. In that way, her opinion would have been supported more strongly. The same speaker failed to elaborate by giving examples when she explained why English is important at Time 3 (Excerpt 12). She said "*because the earth is becoming so globalization*" (line 2). If Student 6 had had provided more examples of what she meant by globalization such as workplace, traveling or education, she might have been more convincing and could have more clearly expressed in which situation English skills are important.

Excerpt (12) Student 6: Comparison Group at Time 3

1. I think :: learning English is important for me :: because company needs person's TOEIC high score and English grade English Eiken high grade
2. Because the earth is becoming so globalization so you should have more English skills right now

In contrast to Student 6, Student 7 from the teacher and peer group successfully expressed her ideas clearly by stating a specific example of why she thinks English is necessary (excerpt 13). Student 7 organized her speech clearly by giving opinions, reasons, and examples. She gave two examples of situations in which English is useful: getting a job (line 3, excerpt 13) and helping foreign tourists (line 6, excerpt 13).

Excerpt (13) Student 7: Teacher and Peer Group at Time 3

1. *In my opinion* learning English is important for me
2. *It's mainly because* :: learning English is useful in the future
3. *For example* when I want :: to work :: many company needs to can speak English
4. So learning English is useful for me
5. And learning English is fun for me
6. *For example* when I go out the city, for example Tokyo
7. Tokyo is many foreigner
8. So *if* foreigner ask me the question :: so *if* I can speak English :: I teach the street or many thing
9. So it is very fun for me
10. And study abroad is a good idea for university students :: because study abroad is good experience
11. *For example* *if* I go study abroad :: I can learn their nature English

The teacher and peer group were trained to produce longer monologues in the following manner: state an opinion followed by reasons or by examples to support the opinion and to elaborate on the idea. In excerpts 13 and 14, the participants successfully elaborated their monologue by stating an opinion first and then supporting the opinion with reasons.

Student 8 was able to give numerous examples for why globalization brings more chances to use English. This participant also gave multiple examples to support her ideas effectively. Excerpt (14) shows that how Student 8 elaborated using examples.

Excerpt (14) Student 8: Teacher and Peer Group at Time 3

1. *In my opinion* I think :: learning English is important for me
2. *It's mainly because* :: it is very useful

3. *For example* when I was high school student :: I went to Hawaii on school trip
4. I went to Hawaii's university and meet with the students
5. He is American
6. So he speak English
7. But I can't communicate with him speedy
8. So it is very sad
9. And I want :: to speak English quickly
10. And I think :: studying abroad is a good idea for university students
11. *It's mainly because* :: I can learn many things
12. *For example* I like western music and culture of foreign countries
13. So I want :: to go abroad and study many things
14. *One reason is* :: I can learn culture of foreign countries
15. So I also learn value of the people
16. *Another reason is* :: I can learn how difficult I speak English
17. *For example* when I am in discussion class :: I can speak English well so

First, Student 8 gave examples of her school trip experience in Hawaii (line 3, excerpt 14), which allowed her to personalize why she thinks English is important. Another example concerned culture; "*I can learn many things*" (line 11) by specifying "*many things*" as music and culture. The last example was from her experience in the discussion class. She was able to convey her messages successfully by talking about personal experiences.

Expressing Possibility

Table 49 shows the descriptive statistics of the target formulaic language that the participants used to express possibility at Time 1, Time 2, and Time 3. Few participants used *if* frequently at Time 1. Both the comparison group and the teacher and peer group increased their use of *if* at Time 2; both groups used *if* 15 times (6 users in the comparison groups *M* = 1.15, seven users in the teacher and peer group *M* = .71), whereas two speakers in the teacher-led group used *if* only four times (*M* = .29). All groups increased their use of *if* at Time 3 (e.g., comparison group nine users, *M* = 1.38, teacher and peer group 14 users, *M* = 1.24).

Table 49. *Descriptive Statistics of Types of Formulaic Language Used for Possibility Function*

		Possibility (If….)		
Group		Time 1	Time 2	Time 3
Comparison (*n* = 12)	Users	3	6	9
	Counts	5	15	18
	M	0.38	1.15	1.38
Teacher-led (*n* = 13)	Users	2	2	7
	Counts	2	4	10
	M	0.14	0.29	0.71
Teacher and peer (*n* = 21)	Users	4	7	14
	Counts	5	15	26
	M	0.24	0.71	1.24

Note. Users = the number of users who used the formulaic language. Counts = frequency of the formulaic language occurrences. *M* = total count/number of participants in the group.

There were few differences between the groups in terms of using *if*. The comparison group used *if* as frequently as the teacher and peer group, possibly because it was more related to the topic compared to the other formulaic language. Another possible reason is that the participants know *if* and it is a high-frequency word that the pedagogical intervention was unlikely to influence their use of *if* for expressing possibility.

The following transcript shows that Student 9 used *if* (line 9, 12) to express possible situations she had never experienced.

Excerpt (15) Student 9: Comparison Group at Time 3

1. I think :: learning English is very important for me :: because recently

2. I can't speak English :: we can't work many company
3. So I think :: it is very important
4. And for example when I eat lunch with my friends :: I talked exchange students in English
5. Then I very surprised :: and I can't speak English well
6. And I can't :: what she speak
7. So I think :: I want :: to understand what she speak
8. I have to study more in English
9. *if* I can understand :: we enjoy speaking with exchange students
10. And I want :: to study English more
11. And go abroad to study is very important :: because the best way :: to improve our English is speaking with other country students
12. *If* we speak with other country :: students our pronounce is improved

The teacher and peer group borrowed the similar usage of *if* from the teacher-modeled passage. Student 9 from the teacher and peer group used *if* five times to state possible situations of going abroad and speaking English better. In line 2, he used *it is mainly because…*, and *if.* The possibility of combining *because* and *if* was demonstrated in the teacher's model during the treatment phase. Given that the frequency of using *if* did not differ between groups, its usage combined with giving a reason was a relatively complicated construction.

Excerpt (16) Student 9: Teacher and Peer Group at Time 3

1. *In my opinion* learning English is important for me
2. *It is mainly because* :: *if* I learn English :: I can enjoying speak a lot of person all over the world
3. It is important for us
4. So *if* I learn English :: I can know the thinking all of the world
5. It is important for all of us
6. *Another reason is* :: I can speak all over the world
7. So and I can make friends all over the world
8. It is also important for us :: to study abroad
9. It is mainly because :: *if* I go abroad :: I can speak only English
10. So English skill will improve too much
11. And *if* we go abroad :: to study :: we can also learn the the country's culture

12. Learning the country's culture is also important for us :: to learn the country
13. And another way :: to stretch English skill is going to English speaking schools
14. English is most important for me

In sum, analyzing the frequency and the variety of target formulaic language showed that both the teacher-led and the teacher and peer groups used the target formulaic language more frequently than the comparison group. The comparison group used non-target phrases such as *I think* or *because* repeatedly to give opinions and reasons.

Second, the participants in the teacher-led and the teacher and peer groups organized their monologues by giving opinions that were supported by reasons and examples compared to their monologues at Time 1. By doing so, they successfully elaborated on their ideas by providing more detailed information. They were able to express their experiences by providing supporting examples for their opinions.

Lastly, the transcriptions also revealed that the degree to which the participants in the experimental groups produced a variety of the target formulaic language differed. The teacher and peer group was able to produce a wider variety of the target formulaic language than the teacher-led group. For example, the teacher and peer group was able to express reasons by saying *It's mainly because….*, *One reason is…*, and *Another reason is…* In contrast, the teacher-led group used one phrase, *It's mainly because….*, to give reasons. This result indicates that the teacher and peer group had been exposed or pressured to use a greater variety of phrases through the peer check activity.

Research Question 4: Communicative Adequacy

Research Question 4 asked about the extent to which the participants who received a form-focused pedagogic intervention developed their communicative adequacy over the 13 weeks. This research question was answered by having 11 raters rate the participants' two-minute oral performances at Time 1, Time 2, and Time 3 using a 5-point rating scale (See Table 9 for the rubric).

The raw scores were statistically analyzed using multifaceted Rasch analysis with the FACETS program version 3.71.4 (Linacre, 2013). The Rasch person ability estimates were examined with a repeated-measures ANOVA in order to examine whether communicative adequacy developed significantly. The independent variables were time (three levels: Time 1, Time 2, and Time 3), and the dependent variable was the Rasch person ability measures.

The infit MNSQ values for raters ranged from 0.7 to 1.3 (Table 50) in the FACETS output so they were well within the acceptable 0.5 to 1.5 range (Fisher, 2007). Initially, the FACETS results showed that Rater 4 misfit with an infit MNSQ statistic of 1.87 and standardized infit statistic of 4.4. A bias analysis was then run using the student-rater-evaluation criteria. The bias analysis indicated the ratings in which Rater 4 had been too lenient or too severe and was inconsistent in his ratings. Rater 4 was inconsistent in his/her ratings. Eight out of 80 misfitting ratings by Rater 4 were replaced with an asterisk in the command file, which meant that those eight ratings were treated as missing data. The analysis was run again and Rater 4 fit the model satisfactorily with an infit MNSQ statistic of 1.42 and a standardized infit statistic of 2.2. Therefore, all raters met the infit MNSQ criterion of .50-1.50.

Table 50 shows the Rasch statistics for the raters for communicative adequacy. The FACETS analysis indicated that the mean Rasch difficulty estimates for the 11 raters ranged from -1.72 to 2.26; Rater 3 had the highest severity estimate followed by Raters 7, 6, 1, 5, 10, 4, 9, 11, 8, and 2, as shown in Table 50; thus, Rater 3 was the most severe rater and Rater 2 was the most lenient. The Rasch reliability estimate for the raters was .97. There were 1,458 inter-rater agreement opportunities and 592 (40.6%) exact agreements. The expected number of exact agreements was 601.7.

Table 50. *Rasch Statistics for the Raters for the Communicative Adequacy*

Rater	Measure	*SE*	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt-measure correlation
3	2.26	.19	0.62	-2.75	0.60	-2.90	.54
7	0.33	.18	0.93	-0.38	0.95	-0.27	.74
6	-0.21	.12	0.96	-0.34	0.97	-0.27	.62
1	-0.35	.18	0.91	-0.58	0.96	-0.20	.57
5	-0.50	.12	0.82	-1.72	0.82	-1.75	.81
10	-0.54	.07	1.09	1.62	1.09	1.60	.62
4	-0.59	.19	1.42	2.26	1.43	2.32	.64
9	-0.69	.18	1.28	1.70	1.28	1.72	.56
11	-0.85	.15	0.94	-0.44	0.94	-0.41	.67
8	-1.55	.17	0.78	-1.50	0.78	-1.47	.52
2	-1.72	.18	0.91	-0.58	0.92	-0.46	.72

Table 51 shows the task measurement report. All of the tasks met the infit MNSQ criterion of .50-1.50, and the standardized statistics did not exceed the ± 2.00 criterion.

Table 51. *Rasch Statistics for the Task Measurement Report*

Time and task	Measure	*SE*	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt-measure correlation
1: Club activity	.00	.07	.99	-.10	1.00	-.01	.67
2: Eating	.00	.07	1.06	1.00	1.05	.88	.65
3: Studying English	.00	.07	.94	-.90	.94	-.91	.71

Note. The three tasks were anchored at 0 logits.

When initially running FACETS, a subset problem arose. Subsets in the data indicate a lack of identifiability of the estimates. This problem impacts the relationship of the measures in the subsets, meaning that it is not possible to compare the measures of the subsets on the logit scale (McNamara, Knoch, & Fan, 2019). One way of dealing with the subset issue is to make some assumptions about the data set. For example, researchers can make the decision that the three task topics are similar in difficulty and specify this in the FACETS software by anchoring the three tasks at the same difficulty level (McNamara et al., 2019). This approach was used in this case; the three tasks were anchored at zero logits.

The mean Rasch item difficulty estimates for each rating component ranged from -.35 to .63 (see Table 52). Fluency had the highest difficulty estimate followed by complexity, organization, and accuracy; thus, fluency was the most difficult criterion and accuracy was the easiest criterion on which to get a high score. Three items, fluency, organization, and accuracy met the infit MNSQ criterion of .50-1.50, and their standardized statistics did not exceed the ± 2.00 criterion. However, the complexity component exceeded the standardized infit criterion of 2.0, as the standardized fit statistic for complexity was -4.7. The part-measure correlation shows the extent to which each component correlated with the total score. The correlation coefficients for fluency, complexity, and organization were similar with values between .70-73. On the other hand, the part-measure correlation for accuracy was smaller (*r* = .59), which suggested that accuracy did not correlate with the total score as much as the other components. The Rasch item reliability estimate of the four components was .95.

Table 52. *Rasch Statistics for the Communicative Adequacy Components*

Rating criterion	Measure	*SE*	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt-measure correlation
Fluency	.63	.08	1.02	0.2	1.02	0.2	.70
Complexity	-.06	.08	0.70	-4.8	0.70	-4.7	.73
Organization	-.22	.08	1.15	2.0	1.14	1.9	.71
Accuracy	-.35	.08	1.13	1.8	1.14	1.8	.59

Table 53 shows the Rasch rating category statistics for communicative adequacy for the two-minute monologue. All categories functioned well according to the rating scale diagnostic criteria: category frequency, average measures, threshold estimates, category fit, and probability curves. Almost half participants' speeches were rated as moderately successful (683 counts, 43%) and 5% (77 counts) were rated as very successful and 2% (30 counts) were rated as unsuccessful.

Table 53. *Ratio of Rating Category for the Communicative Adequacy*

Rating category	Count (%)	Average measure	Outfit MNSQ	Rasch-Andrich threshold	*SE*
5 Very successful	77 (5%)	2.69	1.0	3.78	.13
4 Successful	444 (28%)	1.46	1.0	1.39	.07
3 Moderately successful	683 (43%)	0.40	1.0	-0.99	.07
2 Poor	345 (24%)	-1.08	1.0	-4.18	.20
1 Unsuccessful	30 (2%)	-2.23	1.1	—	—

The FACETS map in Figure 13 represents an overview of the rating results. All facets were measured in uniform units (*logits*), which are indicated on the left side of the map in the measure column. The second column shows the participants' Rasch ability estimates. More competent participants are placed toward the top and less competent participants toward the bottom. The third column shows the rater severity estimates.

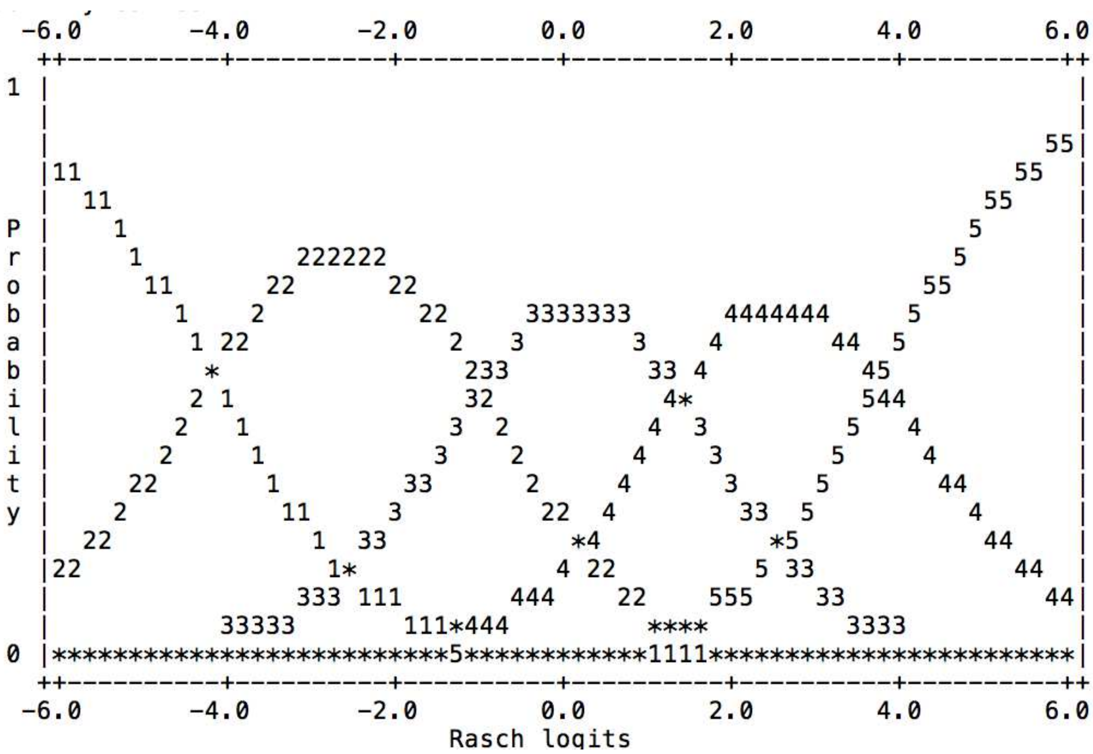
```
------------------------------------------------------------------------------------
|Measure|+Students         |-Raters |-Tasks(CAF)         |-Item        | -Scale    |
------------------------------------------------------------------------------------
|   4   +                  +        +                    +             +   (5)     |
|       |                  |        |                    |             |           |
|       |                  |        |                    |             |           |
|       |                  |        |                    |             |           |
|       | .                |        |                    |             |           |
|       | *                |        |                    |             |           |
|   3   +                  +        +                    +             +           |
|       |                  |        |                    |             |           |
|       | *                |        |                    |             |   4       |
|       |                  |        |                    |             |           |
|       | .                | 3      |                    |             |           |
|       | *                |        |                    |             |           |
|   2   + *                +        +                    +             +           |
|       | *                |        |                    |             |           |
|       | ***              |        |                    |             |           |
|       | **               |        |                    |             |           |
|       | .                |        |                    |             |  ---      |
|       | **.              |        |                    |             |           |
|   1   + ****             +        +                    +             +           |
|       | ****.            |        |                    |             |           |
|       | ***              |        |                    |Fluency      |           |
|       | ****             |        |                    |             |           |
|       | **.              |7       |                    |             |           |
|       | *****.           |        |                    |             |   3       |
*   0   * ***              *        * 1   2   3          *Complexity   *           *
|       | ***.             |6       |                    |Organization |           |
|       | .                |1       |                    |Accuracy     |           |
|       | **               |10 5    |                    |             |           |
|       | ***              |4  9    |                    |             |           |
|       | ***.             |11      |                    |             |           |
|  -1   + ****.            +        +                    +             +           |
|       | **.              |        |                    |             |  ---      |
|       | *.               |        |                    |             |           |
|       | *                |8       |                    |             |           |
|       | *                |2       |                    |             |           |
|       | .                |        |                    |             |           |
|  -2   + **               +        +                    +             +           |
|       | .                |        |                    |             |           |
|       | *                |        |                    |             |           |
|       |                  |        |                    |             |           |
|       | .                |        |                    |             |   2       |
|       | *                |        |                    |             |           |
|  -3   +                  +        +                    +             +           |
|       | .                |        |                    |             |           |
|       | .                |        |                    |             |           |
|       | *                |        |                    |             |           |
|       |                  |        |                    |             |           |
|       |                  |        |                    |             |           |
|  -4   +                  +        +                    +             +   (1)     |
|       |                  |        |                    |             |           |
------------------------------------------------------------------------------------
```

Figure 13. FACET summary for the monologue task.

More severe raters appear toward the top, and more lenient raters appear toward the bottom. Rater 3 was the strictest rater while Rater 2 was the most lenient. The fourth column shows the task difficulty estimate. All the tasks (Club activity = Time 1, Eating =

Time 2, Studying English = Time 3) were at the same level because the three tasks were anchored at 0 logits. The last column shows the difficulty of the four rating categories: organization, accuracy, fluency, and complexity. Fluency was the most difficult, followed by complexity, organization, and accuracy.

The probability curves for the rating categories are shown in Figure 14. The shape of the probability curves for each category formed a peak; thus, each category was clearly distinguished from adjacent categories.



```
   -6.0       -4.0       -2.0        0.0        2.0        4.0        6.0
    ++----------+----------+----------+----------+----------+----------++
  1 |                                                                   |
    |                                                                   |
    |                                                                 55|
    |11                                                             55  |
    |  11                                                         55    |
  P |    1                                                       5      |
  r |     1          222222                                     5       |
  o |      11      22      22                                 55        |
  b |        1    2          22     3333333      4444444     5          |
  a |         1 22             2   3       3    4       44  5           |
  b |          *                233         33 4          45            |
  i |         2 1               32           4*           544           |
  l |        2   1             3  2         4  3         5   4          |
  i |       2     1           3    2       4    3       5     4         |
  t |     22       1        33      2     4      3     5       44       |
  y |    2          11     3         22  4        33  5          4      |
    |  22             1  33            *4           *5            44    |
    |22                1*             4 22         5 33             44  |
    |                333 111       444    22    555    33             44|
    |           33333       111*444         ****         3333           |
  0 |**************************5************1111************************|
    ++----------+----------+----------+----------+----------+----------++
   -6.0       -4.0       -2.0        0.0        2.0        4.0        6.0
                               Rasch logits
```

Figure 14. Probability curves of the rating category.

To address Research Question 4, three repeated-measures ANOVAs were conducted to evaluate the extent to which the comparison group, the teacher-led group, and the teacher and peer treatment group improved on communicative adequacy across the 13 weeks. The independent variable was group (three levels: comparison, teacher-led, and teacher and peer groups), and the dependent variable was communicative adequacy. Repeated-measures ANOVAs were run separately for the three groups. The first set concerned differences at Time 1, Time 2, and Time 3 for the comparison group, the second set was focused on changes at Time 1, Time 2, and Time 3 for the teacher-led group, and the third set concerned differences at Time 1, Time 2, and Time 3 for the teacher and peer group.

Table 54 shows the descriptive statistics for the FACETS measures of communicative adequacy for each group at Time 1, Time 2, and Time 3. All three groups had the highest mean scores at Time 3. The comparison group had a mean score of -0.13 at Time 1, 0.30 at Time 2, and 1.13 at Time 3. They displayed linear development throughout the treatment phase. The teacher-led group had a mean score of -0.85 at Time 1, -0.77 at Time 2, and 0.66 at Time 3. This group did not improve significantly between Time 1 and Time 2 because the mean difference was only .08; however, they improved at Time 3. The teacher and peer group had a mean logit measure of -0.56 at Time 1, -0.38 at Time 2, and 0.78 at Time 3; and they improved by 1.16 logits from Time 2 to Time 3.

Before conducting the repeated-measures ANOVAs, the assumptions of the analysis were checked. First, univariate outliers were checked by converting the raw scores for the speech data to *z*-scores and checking for values > ± 3.29. Second, normality was checked as shown by *z*-skewness and *z*-kurtosis. Normality were met with *z*-skewness and *z*-kurtosis statistics < |2.58|.

The alpha level for the three repeated-measures ANOVAs was set at *p* = .016 (.05/3) in order to avoid committing a Type I error. The first repeated-measures ANOVA was run with data from the comparison group. Mauchly's test indicated that the assumption of sphericity was met, chi-square = 1.10, *p* = .58. The ANOVA was significant, *F*(2, 24) = 9.82, *p* = .001, partial eta square = .45, so three paired samples *t*-tests were run to make post hoc comparisons among time. No significant difference was found between Time 1 (*M* = -.13, *SD* = .42) and Time 2 (*M* = .30, *SD* = .31); *t*(12) = - 1.30, *p* = .22, Cohen's *d* = 1.18, however, there was a significant difference between Time 1 (*M* = -.13, *SD* = .42) and Time 3 (*M* = 1.13, *SD* = .23); *t*(12) = -4.76, *p* < .001, Cohen's *d* = 3.88, and between Time 2 (*M* = .30, *SD* = .31) and Time 3 (*M* = 1.13, *SD* = .23); *t*(12) = -3.11, *p* = .009, Cohen's *d* = 3.07. Thus, the participants in the comparison group improved communicative adequacy significantly at Time 3 compared with Times 1 and 2.

The next repeated-measures ANOVA was run with the teacher-led group. Mauchly's test indicated that the assumption of sphericity was met, chi-square = 3.37, *p* = .19. There was a significant time effect for the teacher-led group, *F*(2, 26) = 3.49, *p* = .001, partial eta square = .43. A paired-samples *t*-test indicated no significant difference between Time 1 (*M* = -.85, *SD* = 1.34) and Time 2 (*M* = -.77, *SD* = .90); *t*(13) = -.24, *p* = .81, Cohen's *d* = 0.16, but there was a significant difference between Time 1 (*M* = -.85, *SD* = 1.34) and Time 3 (*M* = .66, *SD* = 1.56); *t*(13) = -3.22, *p* = .007, Cohen's *d* = 1.82 and between Time 2 (*M* = -.77, *SD* = .90) and Time 3 (*M* = .66, *SD* = 1.56); *t*(13) = -4.12, *p* = .001, Cohen's *d* = 1.16. These results indicated that the participants in the teacher-led group decreased communicative adequacy slightly at Time 2, but improved significantly in the second half of the semester.

Table 54. *Descriptive Statistics for Communicative Adequacy at Times 1, 2, and 3*

	Comparison group			Teacher-led group			Teacher and peer group		
	Time 1	Time 2	Time 3	Time 1	Time 2	Time 3	Time 1	Time 2	Time 3
Minimum	-1.97	-1.07	-0.01	-3.58	-2.75	-3.38	-3.55	-3.13	-2.65
Maximum	3.24	2.10	2.70	1.14	0.77	3.31	2.26	1.59	3.10
M	-0.13	0.30	1.13	-0.85	-0.77	0.66	-0.56	-0.38	0.78
95% CI	[1.04, .77]	[-1.62, -.007]	[.63, 1.62]	[-1.62, -.007]	[-1.29, -.25]	[-.24, 1.56]	[-.37, .97]	[-1.02, .25]	[.23, 1.33]
SD	1.49	1.11	0.81	1.34	0.90	1.56	1.17	1.39	1.22
Skewness	0.89	0.14	0.59	-0.30	-0.53	-1.21	-0.18	-0.52	-0.70
SES	0.62	0.62	0.62	0.60	0.60	0.60	0.50	0.50	0.50
z-skewness	1.45	0.22	0.95	-0.50	-0.88	-2.03	-0.36	-1.05	-1.40
Kurtosis	0.74	-1.40	-0.45	-0.28	0.82	3.07	2.27	-0.79	2.32
SEK	1.19	1.19	1.19	1.15	1.15	1.15	0.97	0.97	0.97
z-kurtosis	0.62	-1.17	-0.38	-0.25	0.71	2.66	2.33	-0.82	2.39

**SES* = Std. Error Skewness, SEK = Std. Error kurtosis.

The third repeated-measures ANOVA was run with the teacher and peer group. Mauchly's test indicated that the assumption of sphericity was met, chi-square = .83, *p* = .66. There was a significant time effect for the teacher and peer group, *F*(2, 40) = 10.38, *p* < .001, partial eta square = .34. Three paired-samples *t*-tests indicated that there was no significant difference between Time 1 (*M* = -.56, *SD* = .25) and Time 2 (*M* = -.38, *SD* = .30); *t*(20) = -.60, *p* = .55, Cohen's *d* = .66; however, there was a significant difference between Time 1 (*M* = -.56, *SD* = .25) and Time 3 (*M* = .78, *SD* = .27); *t*(20) = -4.22, *p* < .001, Cohen's *d* = 5.15, and between Time 2 (*M* = -.38, *SD* = .30) and Time 3 (*M* = .78, *SD* = .27); *t*(20) = -3.35, *p* = .003, Cohen's *d* = 4.07.

Figure 15 shows the changes in communicative adequacy for the comparison, the teacher-led, and the teacher and peer groups. In sum, all groups significantly improved communicative adequacy from Time 1 to Time 3. All three groups improved somewhat from Time 1 and Time 2. However, they improved significantly from Time 2 and Time 3. This result suggests that the participants acquired the ability to produce more effective monologues after Time 2.

The results of this study indicate that all groups developed greater communicative adequacy. The teacher and peer group had the largest effect size (Cohen's *d* = 5.15) between Time 1 and 3. To understand this finding in more detail, I present the raw scores by the human raters for each rating category: organization, syntactic complexity, morphosyntactic accuracy, and oral fluency. In this way I can interpret the development of communicative adequacy by comparing it with the development of the analytical CALF measures.

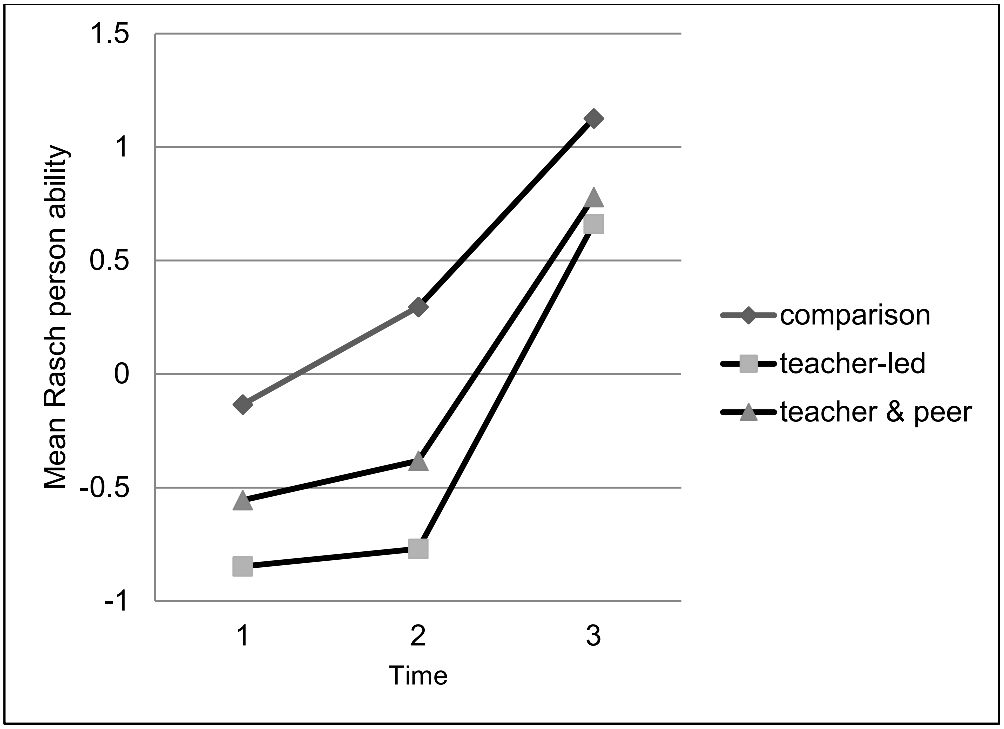


Figure 15. Changes in communicative adequacy at Times 1, 2, and 3.

First, the teacher and peer group's mean scores for organization were 2.97, 3.07, 3.68 at Times 1, 2, and 3, respectively (Table 55); thus, the teacher and peer group gained 24% on the organization variable between Time 1 and Time 3. Similarly, the teacher-led group had a 25% gain between Time 1 and Time 3. Although the comparison group had the highest mean score at Time 1 (*M* = 3.13, *SD* = .86) and remained the highest at Time 3 (*M* = 3.69, *SD* = .82), their gain was 18% between Time 1 and Time 3. This finding suggests that the two experimental groups made more improvement than the comparison group for organization.

The human raters' perceptions of syntactic complexity did not reflect the two analytical complexity measures, mean length of AS units and clauses per AS unit. The findings for morphosyntactic complexity in research question 1 showed that the teacher and peer groups did not significantly improve syntactic complexity, while the comparison group improved both clauses per AS unit and mean length of AS units and the teacher-led group improved clauses per AS unit. In terms of the human raters, the comparison group gained 12% for complexity from Time 1 (*M* = 3.15, *SD* = .78) to Time 3 (*M* = 3.53, *SD*= .76). The teacher-led group gained 22% from Time 1 (*M* = 2.86, *SD* = .65) to Time 3 (*M* = 3.42, *SD* = .83), and the teacher and peer group gained 15% from Time 1 (*M* = 2.90, *SD* = .71) to Time 3 (*M* = 3.44, *SD* = .71) (Table 56). Thus, the two experimental groups improved more than the comparison group, which suggests that the raters' perceptions of complexity differed from the analytical measures of subordinate or AS unit length.

The human ratings for morphosyntactic accuracy were similar to the analytical measures for morphosyntactic accuracy (Table 57). For example, the analytical measures significantly declined between Time 1 and Time 3 for the comparison group. The human ratings for morphosyntactic accuracy improved only 3% from Time 1 (*M* = 3.21, *SD* = .86) and Time 3 (*M* = 3.31, *SD* = .82). The two experimental groups slightly improved morphosyntactic accuracy from Time 1 to Time 3; the gain ratio was 12% for the teacher-led group and 4% for the teacher and peer group. This finding suggests that the participants' improvements were very small from the human raters' point of view as well.

Table 55. *Human Raters' Perceived Organization*

	Comparison group			Teacher-led group			Teacher and peer group		
	Time 1	Time 2	Time 3	Time 1	Time 2	Time 3	Time 1	Time 2	Time 3
Minimum	2	2	2	2	1	1	1	1	1
Maximum	5	5	5	4	5	5	5	5	5
M	3.13	3.27	3.69	2.80	2.80	3.50	2.97	3.07	3.68
SD	0.86	1.10	0.82	0.72	0.82	1.08	0.80	0.98	0.95

One noteworthy finding is that the human raters' perceptions of the participants' fluency somewhat reflected the analytical measures of oral fluency development (Table 58). The comparison group did not improve any analytical measures of fluency, while the teacher-led group significantly improved mean length of run, and the teacher and peer group improved mean length of run and phonation time ratio. The comparison group improved perceived fluency by 12% from Time 1 (*M* = 3.05, *SD* = .89) to Time 3 (*M* = 3.41, *SD* = .84). The teacher-led group improved 28% from Time 1 (*M* = 2.49, *SD* = .85) to Time 3 (*M* = 3.18, *SD* = 1.01), and the teacher and peer group improved 31% from Time 1 (*M* = 2.38, *SD* = .74) to Time 3 (*M* = 3.12, *SD* = .83). Two experimental groups made larger gains than the comparison group. The human raters' impression of the participants' fluency was similar to the analytical measures of fluency.

In summary, all three groups improved communicative adequacy significantly as indicated by the Rasch person ability logits. The teacher and peer group had a large gain in communicative adequacy considering the magnitude of the effect size (Cohen's *d* = 5.15). The percentage of raw score gains for each rating category indicated that the teacher and peer group gained 31% for perceived fluency and 24% for organization over the academic semester. Table 59 shows the summary of percentage gains for the four rating criteria.

Table 56. *Human Raters' Perceived Complexity*

	Comparison group			Teacher-led group			Teacher and peer group		
	Time 1	Time 2	Time 3	Time 1	Time 2	Time 3	Time 1	Time 2	Time 3
Minimum	2.00	2.00	2.00	2.00	1.00	2.00	1.00	1.00	2.00
Maximum	5.00	5.00	5.00	4.00	4.00	5.00	4.00	5.00	5.00
M	3.15	3.32	3.53	2.86	2.80	3.42	2.90	2.98	3.44
SD	0.78	0.71	0.76	0.65	0.72	0.83	0.71	0.80	0.71

Table 57. *Human Raters' Perceived Morphosyntactic Accuracy*

	Comparison group			Teacher-led group			Teacher and peer group		
	Time 1	Time 2	Time 3	Time 1	Time 2	Time 3	Time 1	Time 2	Time 3
Minimum	2.00	2.00	2.00	1.00	2.00	2.00	1.00	2.00	2.00
Maximum	5.00	5.00	5.00	5.00	4.00	5.00	5.00	5.00	5.00
M	3.21	3.41	3.31	3.03	3.08	3.39	3.27	3.20	3.39
SD	0.86	0.69	0.82	0.82	0.80	0.86	0.88	0.83	0.77

Table 58. *Human Raters' Perceived Fluency*

	Comparison group			Teacher-led group			Teacher and peer group		
	Time 1	Time 2	Time 3	Time 1	Time 2	Time 3	Time 1	Time 2	Time 3
Minimum	2.00	2.00	2.00	1.00	1.00	1.00	1.00	1.00	1.00
Maximum	5.00	5.00	5.00	4.00	5.00	5.00	4.00	4.00	5.00
M	3.05	3.30	3.41	2.49	2.73	3.18	2.38	2.73	3.12
SD	0.89	0.81	0.84	0.85	0.93	1.01	0.74	0.81	0.83

Table 59. *Percentage Gains from Time 1 to Time 3 for the FACETS Rating Criteria*

	Comparison group	Teacher-led group	Teacher and peer group
Organization	18%	25%	24%
Complexity	12%	22%	15%
Accuracy	3%	12%	4%
Fluency	12%	28%	31%
Effect size	3.88	1.82	5.15

Note. Cohen's *d* = 3.88 (Comparison group); 1.16(Teacher-led group); 5.15(Teacher & peer group).

Research Question 5: The Relationship Between CALF and Communicative Adequacy

Research Question 5 asked about the relationship between the analytical CALF measures and the human ratings of communicative adequacy. This research question was answered by conducting a Pearson correlation analysis. The four criteria—Organization, Complexity, Accuracy, and Fluency—were used to compute estimates of communicative adequacy. The computed value was the communicative adequacy score measured using Rasch logits as shown in Table 54.

Pearson correlation coefficients were computed among the CALF measures and communicative adequacy. The descriptive statistics are shown in Table 60. All the values are combined from the comparison group, the teacher-led group, and the teacher and peer group because the purpose of this correlation analysis was to assess the relationships among the CALF measures and communicative adequacy, not to compare the groups. The communicative adequacy measure was the lowest at Time 1 (*M* = -0.53) and the highest at Time 3 (*M* = .84). Other measures were also at the highest at Time 3.

Table 60. *Descriptive Statistics of Communicative Adequacy and CALF Measures*

	Time 1		Time 2		Time 3	
	M	*SD*	*M*	*SD*	*M*	*SD*
Communicative adequacy	-.53	1.31	-.31	1.24	.84	1.23
Clauses per AS unit	1.65	.29	1.58	.31	1.90	.36
Mean length of AS units	10.55	1.73	9.82	1.85	11.90	2.16
Error Free AS units	.66	.15	.62	.21	.65	.18
MTLD	36.66	9.50	31.08	6.83	43.68	10.46
Mean length of pauses	.67	.18	.72	.20	1.36	.42
Number of repair occurrences	9.73	6.71	9.92	6.20	9.79	4.95
Mean duration of syllable	.29	.04	.32	.03	.30	.03
Mean length of run	4.38	.83	4.41	.97	5.16	.86
Phonation time ratio	49.78	8.66	50.65	9.42	55.38	8.32

Note. MTLD = *Measure of Textual Lexical Diversity*.

Prior to conducting correlational analysis, assumption of normality was checked. As shown previously, variables are normally distributed except for one variable clauses per AS unit for Time 2. Therefore, correlation coefficient at Time 2 should be interpreted with caution.

Table 61 shows the results of the correlational analyses between the Rasch measures of communicative adequacy and the CALF measures at Times 1, 2, and 3. Plonsky and Oswald (2014) suggested that *r* close to .25 is a small effect, .40 is a medium effect, and .60 is a large effect in the field of L2 research.

Syntactic complexity had little relationship with communicative adequacy. Both clauses per AS units and mean length of run showed small correlations with communicative adequacy (r = .14–.28, p > .05). This finding suggests that the participants' production of more complex and longer utterances had little to do with their perceived performances by human raters.

Table 61. *Correlations Among Communicative Adequacy and the CALF Measures*

	Time 1	Time 2	Time 3
Complexity			
Clauses per AS unit	.14	.28	.19
Mean length of AS units	.13	.13	.25
Accuracy			
Error Free AS unit	.29*	.09	.12
Lexis			
MTLD	.13	.35*	.02
Fluency			
Mean length of pauses	.40**	.51**	-.63**
Number of repair occurrences	-.15	.00	.04
Mean duration of syllable	-.13	-.15	-.15
Mean length of run	.38**	.45**	.33*
Phonation time ratio	.44**	.61**	.60**

Note. ** Correlation is significant at < .01 (2-tailed). * Correlation is significant at < .05 (2-tailed). MTLD = *Measure of Textual Lexical Diversity*.

Error-free AS units, which was a measure of morphosyntactic accuracy, had a weak relationship with communicative adequacy (r = .29, p < .05) at Time 1, in which the participants talked about a club activity. There was almost no relationship between morphosyntactic accuracy and communicative adequacy at Time 2 and Time 3, r = .09 and r = .12, respectively, p > .05. This finding suggests that the participants' grammatical accuracy did not result in higher communicative adequacy scores when they talked about eating out (Time 2) and studying English (Time 3).

MTLD, a measure of lexical diversity, had a weak relationship with communicative adequacy at Time 1 (r = .13, p > .05) and Time 3 (r = .02, p > .05), and a medium relationship with communicative adequacy at Time 2 (r = .35, p < .05), in which the participants talked about eating out.

In general, fluency had a more consistent relationship with communicative adequacy than the other CALF measures. For example, mean length of run had a medium relationship with communicative adequacy at Time 1 (r = .38, p < .01), Time 2 (r = .45, p

< .01) and Time 3 (*r* = .33, *p* < .01). This result shows that the longer the run the participants produced, the higher the communicative adequacy scores were regardless of the topic. Phonation time ratio also had a strong relationship with communicative adequacy at Time 1 (*r* = .44, *p* < .01), Time 2 (*r* = .61, *p* < .01), and Time 3 (*r* = .60, *p* < .01); thus, the longer the participants spoke, the higher their communicative adequacy measures were regardless of the topic. Mean duration of syllable had a small and negative relationship with the communicative adequacy scores at Time 1 (*r* = -.13, *p* >.05), Time 2 (*r* = -.15, *p* > .05), and Time 3 (*r* = -.15, *p* >.05). This finding suggests that longer syllables were associated with lower communicative adequacy scores.

Repairs had a weak relationship with communicative adequacy at Time 1 (*r* = -.15, *p* >.05), Time 2 (*r* = .00, *p* > .05), and Time 3 (*r* = .04, *p* >.05); thus, communicative adequacy did not decline even when the participants self-corrected or repeated themselves more frequently.

Mean length of pauses had mixed results. Long pauses usually indicate disfluency. At Time 3, the finding shows that longer pauses were associated with lower communicative adequacy measures (*r* = -.63, *p* > .05). However, mean length of pauses had a positive medium relationship with communicative adequacy at Time 1 (*r* = .40, *p* < .01) and Time 2 (*r* = .51, *p* < .01); thus, longer pauses were associated with higher communicative adequacy. Given that pauses were calculated as a total amount, the total pause time might not have been important if the participants sounded fluent once they spoke.

In sum, fluency had a stronger relationship with communicative adequacy than syntactic complexity, morphosyntactic accuracy, and lexis. Mean Length of run a had medium relationship and phonation time ratio had a strong relationship with communicative adequacy. However, repairs had a weak relationship with communicative adequacy. The findings showed that the length and amount of speaking was somewhat related to communicative adequacy. Although the participants produced many repairs, as long as they spoke smoothly in one run, the human raters judged them as sounding more communicatively effective.

Research Question 6: Learners' Prioritization While Performing the Speaking Task

Research question 6 asked what the learners prioritized during the monologue recording. This research question was answered by administering a retrospective questionnaire. The participants in each group indicated what they prioritized during the monologue immediately after completing the 3/2/1 task at Times 1, 2, and 3 (see Appendix F for the 3/2/1 recording retrospective questionnaire). On the questionnaire, the students indicated what they had prioritized by choosing one or more of the following options: focus on content, organization, grammar, lexis, and formulaic language.

The next coding stage was to categorize the participants' descriptions about each category into more detailed codes by closely examining the participants' comments. The participants' descriptions in each category were sorted with the similar features of the participants' answers adapted from relevant coding which Pang and Skehan (2014) used. Finally, the comments concerning content were placed into three categories: topic relevance, self- relevance, and think of capability. Comments concerning organization were categorized as logical organization, elaboration, or time management. Grammar comments were categorized as general use or specific use. Comments concerning lexical retrieval were categorized as general use and specific use. Target formulaic language had one sub-category, using the target formulaic language (e.g., *In my opinions, One reason is…*). Table 62 shows the coding categories. Content planning is related to Levelt's conceptualizer stage in which speakers process ideas that feed into the pre-verbal massage. Lexical and grammar planning are concerned with Levelt's formulator stage in which speakers transform the preverbal message into a linguistic form.

Table 62. *Coding Scheme: Speakers' Focus While Speaking the 3/2/1 Recording*

CODE	Description of CODE	Example
Content		
Topic relevance	Think of topic related content	I tried to generate ideas to answer the topic.
Self-relevance	Think of experiences	I tried to relate to my experiences.
Capability	Think of own capability	I tried to think what I can say in English. I tried to think if my English make sense to a listener.
Organization		
Logical organization	Try to organize the ideas in a coherent manner	I used *First* or *Second*. I tried to talk in a chronological order.
Topic development	Try to elaborate ideas with examples and reasons	I tried to elaborate using many examples. I said my opinions first then I supported them with reasons.
Time management	Try to manage time	I tried to manage time to finish answering the questions.
Grammar		
General use	Think of grammar in general	I tried to focus on grammar.
Specific use	Think of particular grammar	I tried to focus on verb tense.
Simple use	Think of simple grammar	I tried to use simple grammar.
Lexical		
General retrieval	Try to find appropriate words	I tried to retrieve words.
Specific retrieval	Choose to use specific words	I tried to avoid the same words over and over.
Simple words	Choose to use simple words	I tried to use simple words.
Target formulaic language		
General use	Think of use of the target formulaic language	I tried to use *In my opinion*.

Coding was checked by a research assistant, who coded 10% of the data. The agreement rate between the research assistant and I was 79.0%. Instances of disagreement were discussed and until 100% agreement was achieved. The data for the coded retrospective questionnaires were entered into an Excel file.

Table 63 shows the frequency of each category during the 3/2/1 recording. I compared the frequency within the categories as groups given that the number of participants in each group differed depending on the group.

First, the participants mainly focused on content while they spoke. The number of instances of a focus on content at Times 1, 2, and 3 was 116 (Comparison group = 35, Teacher-led group = 35, Teacher and peer group = 46). The two most frequently selected sub-categories were content-relevance (*n* = 55) and self-relevance (*n* = 40). These two sub-categories were similar because the participants who thought about their own experience also thought about the topic; however, topic relevance was more general, as it included statements such as *I think of general ideas and my ideas* and *I try not to talk off-topic* (一般論と自分の考えについて考えた). In contrast, self-relevance was more specific, as it included comments such as *I tried to talk about the club which I belong to now* （自分のクラブ活動について話そうとした）or *I tried to talk about my study abroad experience* （自分の留学経験について話そうとした）. Other participants focused on content by considering their own speaking proficiency (*n* = 21). In these cases, the participants wrote comments such as *I chose what I can talk easily* （話しやすい内容を選んだ）and *I firstly prioritized what I can easily express*（まず簡単に表現しやすい内容を優先的に話した）. Among the answers concerning speaking capability, two participants answered from a listener's point of view by stating *I think about whether my content is easy for the listener to understand* （内容が相手に理解されやすいかを考えた）and *I tried to think of a topic that anyone can understand*（誰でも理解できるようなトピックを考えるようにした）. Although the recording was completed individually without a partner, these answers indicated that they thought of achieving a communicative goal by attempting to convey their meaning clearly.

Second, some participants prioritized organization. The participants who answered that they focused on organization mostly answered that they thought about how to develop the topic (*n* = 34). Comments included *I tried to use many examples* （例を沢山出そうとした）and *After I said my opinion, I tried to show a clear reason to support my idea* （意見を言った後、理由を言って意見をサポートしようとした）. Some participants (*n* = 15) focused on logical organization. Although topic development and logical organization cannot be separated completely, logical organization is more related to speech structure. Representative comments included *I tried to put things into order* （順序立てて話すようにした）*, I tried to say things chronological order*（時系列に沿って話そうとした）, and *I tried to say first, second …* （一つ目、二つ目、って話そうとした。）A few students wrote that they thought about time management.

Table 63. *Frequency Data for All Codes*

CODE	Comparison group			Teacher-led			Teacher and peer			
	Time 1	Time 2	Time 3	Time 1	Time 2	Time 3	Time 1	Time 2	Time 3	Total
(A) Content										
Topic relevance	3	7	6	3	8	7	7	8	6	55
Self-relevance	6	4	5	4	3	3	6	3	6	40
Capability	2	0	2	6	0	1	5	4	1	21
	11	11	13	13	11	11	18	15	13	116
(B) Organization										
Logical organization	1	1	2	1	1	3	2	2	2	15
Topic development	2	2	0	6	3	4	6	7	4	34
Time management	1	0	2	1	0	0	0	3	2	9
	4	3	4	8	4	7	8	12	8	58
(C) Grammar										
General use	1	0	1	2	0	2	1	1	2	10
Specific use	1	1	5	2	5	2	2	3	4	25
Simple use	1	3	0	3	1	0	3	1	1	13
	3	4	6	7	6	4	6	5	7	48
(D) Lexis										
General retrieval	2	2	1	2	4	1	3	0	4	19
Specific retrieval	2	3	0	4	1	0	2	1	1	14
Simple words	1	3	4	2	1	1	1	0	2	15
	5	8	5	8	6	2	6	1	7	48
(E) Target formulaic language										
General use	N/A	9	7	N/A	7	13	N/A	19	18	73

Kenta from the teacher-led group said in the interview that *We did in a solo-recording situation. Nevertheless, I think conveying the meaning is important. If I just say one sentence, people would not understand me. So, I tried to think of how I organize the monologue* （一応、録音とはいえ、伝えるのが重視なんじゃないかなっていう考え方なんで、一文ベラベラ並べただけじゃ、人ってわかんないじゃないですか、だからなんだろ、構成はわりと気にしているのかな、気にしてたんですね）.

Third, the participants did not prioritize grammar as frequently as content. In total, there were 48 instances of a grammar focus. When the participants prioritized grammar, they mainly focused on grammar in general by trying not to make mistakes (*n* = 10) or by using a specific grammatical feature (e.g., verb tenses) correctly (*n* = 25). In both cases, the participants' reason for focusing on grammar was to pay increased attention to correct grammar usage. On the other hand, 13 students attempted to avoid making grammatical errors by avoiding the use of complex grammar. These participants stated, *I tried to use the junior high school level grammar and vocabulary* （中学生の分方や語彙を使おうとした）or *I focused only on simple grammar and vocabulary* （簡単な文法や語彙だけを使った）. Lexical retrieval was similar to the grammar focus. In total, there were 48 counts of a lexical focus. They mainly focused on general lexical retrieval (*n* = 19) followed by specific retrieval (*n* = 14) such as *I tried to come up with miso-shiru (miso-soup) in English*（お味噌汁ってなんていうのかなって思いました）. The participants also thought about using simple words (*n* = 15).

Lastly, some participants prioritized the use of the target formulaic language (*n* = 73). On the first retrospective questionnaire that was administered after the first recording, one item asking about the target formulaic language was not shown because the participants had not been exposed to the pedagogical intervention at that time. At Time 2 and Time 3, seven to nine participants in the comparison group answered that they focused on the target formulaic language. Given that the participants in the comparison learned the formulaic language outside of the treatment, some of them were able to apply their knowledge during the 3/2/1 task. Seven (Time 2) and 13 (Time 3) out of 14 participants in the teacher-led group prioritized using the target formulaic language. In the teacher and peer group 18 (Time 2) and 19 (Time 3) out of the 21 participants answered that they prioritized the target formulaic language.

In sum, the participants mainly focused on meaning over form. Even when the participants focused on linguistic form, such as grammar or vocabulary, they thought about simple ways to use language. In addition, some of the participants tried to think of organization and formulaic language while engaging in the 3/2/1 task.

CHAPTER 5

DISCUSSION

In this chapter, the results for the six research questions are summarized and interpreted, and at the end of the chapter I present research implications and pedagogical implications with regards to second language oral performance. Data from the interviews with the four students who volunteered to participate as interviewees are included to better explain some of the quantitative findings. Nana was from the comparison group, Kenta was from the teacher-led group, and Megu and Sumi were from the teacher and peer group.

The Development of Speaking Proficiency

Research question 1 concerned the extent to which the comparison group, the teacher-led group, and the teacher and peer group improved syntactic complexity, morphosyntactic accuracy, lexis, and oral fluency in this 13-week study. This research question was designed to address two research gaps. The first gap is that few researchers have explored the longitudinal effects of pre-task planning. Many TBLT researchers (e.g., Foster & Skehan, 1999; Geng & Ferguson, 2013; Ortega, 1999, Tavakoli & Skehan, 2005) have conducted cross-sectional studies in which they found that pre-task planning is effective, especially in terms of developing oral fluency and syntactic complexity compared to a no-planning condition. The second gap is that the longitudinal development of CALF through the 3/2/1 task has rarely been investigated and when it has been investigated, the development of some CALF variables has been ignored. De Jong and Perfetti (2011) investigated the longitudinal effects of the 4/3/2 task and found that repeating the same topic enhanced learners' oral fluency because it led to greater proceduralization of linguistic knowledge compared to using different topics; however, the learners' development in terms of syntactic complexity, morphosyntactic accuracy, and lexical diversity were not reported. This research question was answered by conducting three sets of repeated-measures ANOVAs. The independent variable was time (Three times: Week 2, 8, and 14), and the dependent variables were the nine CALF measures (see Table 7 for CALF measures).

Summary of the Results for the Development of Speaking Proficiency

The results showed the following characteristics of the participants' longitudinal development of speaking proficiency in terms of the CALF measures. First, the development of the participants' speaking proficiency was non-linear between Times 1, 2, and 3 for many of the CALF measures, as the participants' CALF scores sometimes decreased at Time 2 or Time 3. For example, all groups increased mean duration of syllable at Time 2, but the measure decreased at Time 3; thus, the participants became more disfluent at Time 2 and more fluent at Time 3. Another example is that all groups decreased MTLD, the measure of lexical diversity, at Time 2, but they increased at Time 3; this pattern indicated non-linear development.

Summarizing the results for research question 1, the comparison group significantly increased syntactic complexity between Time 1 and Time 3 for clauses per AS unit (p = .001, Cohen's d = 1.27) and mean length of AS units (p < .001, Cohen's d = 1.33). They also significantly increased MTLD between Time 1 and Time 3 (p = .002, Cohen's *d* =1.13). On the other hand, the participants became more disfluent, as they significantly increased mean duration of syllable between Time 1 and Time 2 (p = .004, Cohen's *d* = 1.11) and between Time 1 and Time 3 (p = .009, Cohen's *d* = .66), and mean length of pauses between Time 1 and Time 3 (p < .001, Cohen's *d* = 1.68). In addition, the comparison group accuracy (error free AS-unit) declined throughout the semester.

The teacher-led group significantly increased complexity measures, clauses per AS unit between Time 1 and Time 3 (p = .01, Cohen's *d* = .87). They significantly increased oral fluency as shown by mean length of run between Time 1 and 3 (p = .01, Cohen's *d* = .79) and Time 2 and Time 3 (p = .001, Cohen's *d* = 1.08). At the same time, they became more disfluent as shown by increases in mean length of pauses between Time 1 and Time 3 (p = .001, Cohen's *d* = 2.14) and between Time 2 and Time 3 (p = .001, Cohen's *d* = 2.03).

The teacher and peer group's oral fluency measures showed a similar developmental pattern as the teacher-led group. They significantly increased mean length of run between Time 1 and Time 3 (p < .001, Cohen's *d* = 1.31) and between Time 2 and Time 3 (p = .003, Cohen's *d* = 0.77); they also increased phonation time ratio between Time 1 and Time 3 (p = .003, Cohen's *d* = 0.76). At the same time, they became more disfluent as shown by increases in mean length of pauses between Time 1 and Time 3 (p = .001, Cohen's *d* = 3.54) and between Time 2 and 3 (p < .001, Cohen's *d* = 3.20).

Although the participants in the two experimental groups improved mean length of run, they also produced longer pauses. In order to understand this finding in greater depth, a further analysis was conducted by looking at the gains of the mean length of run and the participants' use of the target formulaic language. The detailed analysis suggested that the top six participants who gained mean length of run the most also produced the target formulaic language more frequently, suggesting that the high mean length of run measures might be associated with a higher frequency of target formulaic language use (see Table 40).

To summarize, the two experimental groups improved mean length of run without sacrificing their performance on measures such as accuracy and lexis. The teacher-led group improved complexity as well. On the other hand, the comparison group improved complexity and lexis but their accuracy decreased and pause length and syllable duration increased; thus, they produced more errors and became less fluent. This finding suggests that L2 learners will not be overwhelmed if they have the chance to engage in task repetition while acquiring and proceduralizing target formulaic language. Table 64 shows a summary of CALF development for each group. Upward arrows indicate an increase of the measure and the downward arrows indicate a decline between Time 1 and Time 3. The increase of pauses and mean duration of syllable suggests disfluency.

Table 64. *Summary of Long-Term Development of CALF between Time 1 and Time 3*

Group	Complexity	Accuracy	Lexis	Fluency
Comparison (*n* = 13)	Clauses/AS ↑ Length/AS ↑	↓	↑	Pauses ↑ MDS ↑
Teacher-led (*n* =14)	Clauses/AS ↑	*ns*	*ns*	Pauses ↑ MLR ↑
Teacher and Peer (*n* = 21)	*ns*	*ns*	*ns*	Pauses ↑ MLR ↑ PTR ↑

Note. ns = not significant. Clauses/AS = Clauses per AS unit; MDS = Mean Duration of Syllable; MLR = Mean length of run; PTR = Phonation Time Ratio. Upward arrows (↑) indicate an increase of the measure from Time 1 to Time 3 and the downward arrows (↓) indicate a decline from Time 1 to Time 3.

Interpretation of the Results for the Development of Speaking Proficiency

First, the results indicated that using text enhancement and the peer-check of formulaic language possibly led to greater oral fluency over the academic semester as assessed by MLR. The teacher-led planning group made a 15.90% gain and the teacher and peer group made a 30.10% gain in MLR, while the comparison group increased only 3%. This result suggests that it is plausible the instruction of formulaic language during the 3/2/1 task might help the learners improve mean length of run, possibly because the speakers were able to access prefabricated chunks stored in long-term memory (Boers & Lindstromberg, 2012; Segalowitz, 2003; Tavakoli et al., 2016; Wood, 2009).

This finding somewhat supports previous studies in which oral fluency development occurred through repetition in one academic semester (e.g., De Jong & Perfetti, 2011; Tavakoli et al., 2016). De Jong and Perfetti explained that repeating the same topic improves oral fluency more effectively than talking about different topics because the greater degree of conceptual and lexical repetition can lead to proceduralization. If this is the case, then it is plausible that L2 speakers can more easily access concepts, lexis, and morphosyntax in the second and third performances because they were activated in the first performance.

Evidence that formulaic sequences help language learners sound more fluent has been reported by Boers et al. (2006) and Wood (2009, 2010). Such fluency increases can occur because ready-made chunks are usually easier and faster to retrieve than generating sentences word by word (Boers et al., 2006). In Wood’s case study (2009), a Japanese female student made a 26.30% gain in mean length of runs between the pretest and posttest after the six-week training session, which included listening to a native-speaker model, drawing attention to formulaic language, a dictogloss task, and a 4/3/2 task. The current study supports the idea that instructional approaches that provide learners with ample output practice to foster the proceduralization of linguistic knowledge possibly hold greater promise for oral fluency development than 3/2/1 speaking practice alone. It might be partly because the repetition and time pressure during the 3/2/1 task helped the participants' proceduralize and perhaps automatize the target formulaic language.

The teacher-modeled passage possibly might have helped the participants allocate attentional resources to monitoring how to express their ideas. By reading the model input, the participants might have noticed some of the target formulaic language because of the underlining cues (Doughty, 1991; Sharwood-smith, 1993). The target formulaic language potentially allowed the speakers to organize their monologues more efficiently by using target phrases such as *in my opinion,* followed by phrases such as *one reason is* and *for example* when elaborating on the initial opinion. This type of organization might allow the participants to understand form-meaning-function mapping because they can understand how and when to use the target form. For instance, Sumi, a participant in the teacher and peer group, said that the formulaic language in the teacher-modeled passage was useful, especially in the beginning of the semester, because she was unsure of how to use the target forms in communicative contexts. She also mentioned that if she understood how to express certain linguistic functions, she could talk more easily. This statement implies that formulaic language functioned as a prompt that encouraged the participants to organize and express their ideas more efficiently.

While input enhancement can raise awareness of noticing the target phrases (Sharwood-Smith, 1993), it might not guarantee that the speakers use them during a speaking task. On the other hand, the peer-check activity might have given the speakers pressure to use the target formulaic language. The peer check function also acted as peer feedback of target form usage, and the speakers knew which of the target formulaic language they did or did not use. Indeed, Megu, who was in the teacher and peer group, stated that she became more motivated to use the target formulaic language because her performances were being checked by a partner. From my observations and the participants' interview data, speakers in the teacher and peer group struggled to attend to both the target formulaic language and the content of their talk in the beginning stages of the 3/2/1 training. This issue likely arose because attentional resources are limited (VanPatten, 1990; Skehan, 1998); therefore, attention cannot be simultaneously placed on linguistic form and meaning. As time went by, the students became more at ease using the target formulaic language during the 3/2/1 task.

The findings of gains in mean length of run can be explained from two theories of automatization. First, in line with Anderson's Adaptive Control of Thought (ACT) theory (1983, 1987), it is plausible that proceduralization and automatization were facilitated in both the teacher-led and the teacher and peer groups through the provision of communicatively oriented declarative knowledge (e.g., text enhancement in a meaningful context) and by having opportunities to repeatedly use the target forms in the 3/2/1 task. Providing form-focused instruction for formulaic language and creating opportunities for practicing the target formulaic language might have facilitated the proceduralization of that language (DeKeyser, 2001, 2007; Segalowitz, 2010). Considering that the target formulaic language was taught explicitly in the initial stages of the study, the repeated and consistent application of the explicitly learned forms during the 3/2/1 task might have created implicit and/or automatized rules. Second, Logan's instance theory also can provide explanation for the acquisition of formulaic language. The participants might have been able to automatically retrieve chunks of the formulaic language as a whole without needing to apply syntactic rules. With experience and practice, the speed of memory retrieval exceeds that of rule-based processing, formulaic language is accessed in memory as one unit (Kormos, 2006, p. 46). In this sense, the participants might have memorized and used certain instances of formulaic language to achieve a particular communicative function. Therefore, the gain in mean length of run could possibly explained by not only the automatic application of rules as described in ACT theory, but also the memory retrieval of appropriate formulaic language as one unit in speech processing as suggested in Logan's instance theory (Kormos, 2006). However, Kormos stated that the automatization of speech production processes is highly speculative and more research is needed to test this possibility (p. 48).

Although the participants in the two experimental groups improved mean length of run, they also produced longer pauses. De Jong and Perfetti (2011) found that when the participants talked about the same topic during the 4/3/2 task for three weeks, mean length of run remained the same while the mean length of pauses decreased. These two measures are key indicators that proceduralization is occurring; thus, if either of these measures improves while the other remains stable, an argument can be made that learners have become more fluent as a result of proceduralization (De Jong & Perfetti, 2011; Doe, 2017; Towell et al., 1996). The participants in both experimental groups in this study increased mean length of run while increasing mean length of pauses, indicating that they paused longer when producing longer runs. Given that Towell et al. (1996) has suggested participants such as those in this study might not have proceduralized their explicit knowledge because their pause length did not decrease, I investigated this issue in greater depth by conducting detailed analyses of the participants' pausing behavior in the 3/2/1 task. The further analyses suggested that there might be an association with gains in the frequency of usage of the target formulaic language and gains in mean length of run (see Table 40).

Given that the participants' MLR gains are related to their use of the target formulaic language, there are two possible reasons why the teacher and peer group increased both mean length of run and mean length of pauses simultaneously. The first reason is that these participants paused to pay attention before using the target formulaic language. For instance, Excerpt 1 shows that Student 6 paused for 1.33 seconds before saying the target form *for example* at Time 3.

Excerpt 1: (student 6, time 3)

> "It's mainly because (0.43) I can learn many things **(1.33) <eh> for example** (1.49) {I} (1.31) I like western music and culture of foreign countries" (Student 6, Time 3).

Student 6 had already given a reason for why she thinks studying abroad is a good idea for university students. She continued to support her opinion by giving an example after pausing. She might have paused to think of the formulaic language and/or an example.

Excerpt 2 shows another example of Student 6's performance at Time 3; she paused 1.60 seconds before using the target phrase *another reason*:

Excerpt 2: (student 6, time 3)

(1.60) <eh> another reason is (0.75) I can learn how difficult (0.76) I speak English.

The second reason for the simultaneous increase is that the participant spent time planning after producing the target formulaic language; thus, the target formulaic language sometimes functioned as a filler to give the speakers time to think about what to say next. For instance, Excerpt 3 shows that Student 5 paused for 3.76 seconds after saying *for example*:

Excerpt 3: (student 5, time 3)

<eh> **for example (3.76)** <eh> {you can} (0.36) {you can} (2.26) <eh> you can see the (0.40) movie by you tube (0.36).

After stating *for example*, Student 5 reformulated his utterance and repeated *you can* three times, which suggests that he accessed both the conceptualizer and the formulator as he was trying to think of both what to say and how to say it. Excerpt 4 shows that Student 1 paused for 1.91 seconds after saying *in my opinion*:

Excerpt 4: (student 1, time 3)

{in my opinion} (1.91) <eh> in my opinion learning English is important for me (0.77)

The speaker had to repair by using self-repetition when he said *in my opinion* twice. Repeating the phrase suggested that he was thinking of what to say. That might be why he also spent time accessing appropriate lexis in order to express himself. In this regard, online planning time, as indicated by longer pauses, is plausibly needed when learners engage in the process of increasing grammatical complexity (e.g., Yuan & Ellis, 2003).

The findings of this study did not show linear development for all the CALF constructs. The participants were not able to use their attentional resources more efficiently and unable to attend to multiple CALF constructs simultaneously. The findings did not support Vercellotti's (2017) examination of longitudinal changes in CALF indices with 63 ESL students in an intensive English program in United States. Vercellotti found that grammatical complexity, accuracy, and fluency showed linear development without any trade-off effect in a study that took place over three academic semesters. One possible reason is the educational setting. Compared to ESL settings, it might take more time for EFL students to develop the CALF constructs because of their relatively limited opportunities to encounter target language input and produce output both inside and outside of classroom. Although more studies need to be conducted in different educational contexts, the current study shed light on the long-term development of CALF measures by EFL learners; the results indicated that low-proficiency Japanese EFL learners cannot improve all CALF areas simultaneously in one academic semester. The second possible reason of not improving all of the CALF indices simultaneously might be because the CALF measures might not be sensitive enough to show the small amounts of development made by low-intermediate L2 learners. For example, it is quite difficult for the L2 learners achieve error-free AS units considering that a minor error that does not compromise comprehensibility, such as omitting third person singular *–s*, is counted as an error. Lastly, the participants in all groups struggled at Time 2, in which they talked about eating in or eating out. Although the topics were selected carefully after conducting the pilot study, the participants experienced some difficulty speaking about that topic, which contributed to the non-linear development some of the CALF measures.

The findings also showed that morphosyntactic accuracy did not develop through the 3/2/1 task regardless of the pedagogical intervention. The comparison group even significantly decreased morphosyntactic accuracy and the two experimental groups did not display significant differences in morphosyntactic accuracy between Time 1 and Time 3. This result is understandable considering that the participants did not have any opportunities to acquire the correct morphosyntactic form through explicit instruction on specific grammatical forms and corrective feedback on grammatical errors. In other words, learners do not acquire new linguistic forms without noticing them and having opportunities to use them communicatively (Schmidt, 1990).

Given that the participants did not improve on all the CALF measures, it is noteworthy to mention that the two experimental groups did not sacrifice other indices such as complexity, accuracy, and lexis. This finding suggests that having an additional pedagogical intervention attached to the 3/2/1 task did not harm other areas of CALF.

Between-Group Performances on the CALF Measures

Research Question 2 concerned the extent to which the teacher-led and the teacher and peer treatment groups outperformed the comparison group in terms of syntactic complexity, morphosyntactic accuracy, lexical diversity, and oral fluency. Previous researchers have examined the effects of the 3/2/1 task (e.g., Boers, 2014; De Jong & Perfetti, 2011; Thai & Boers, 2016), but they did not investigate a pedagogical intervention associated with the task. Examining the statistical differences between groups in terms of the CALF variables shed light on the effects of the two form-focused interventions. This research question was answered by conducting a one-way ANOVA to compare the differences between groups for Time 1 and ANCOVAs for Time 2 and Time 3. The independent variable was group (Three levels: the comparison group, the teacher-led group, and the teacher and peer group), and the dependent variables were the nine CALF measures (Table 7).

Summary of the Results for the Significant Differences Between Groups

The comparison group produced more clauses per AS-unit, longer AS-unit, and shorter pauses compared to the teacher and peer group at Time 3. The comparison group performed better in terms of producing more complex utterances with less pauses at Time 3. The lack of significant group differences for the CALF measures indicated that the pedagogical intervention the two experimental groups received did not greatly affect the participants' oral performances. In this section, I discuss why the experimental groups did not significantly outperform the comparison group in terms of the CALF measures.

Interpretation of the Results for the Significant Differences Between Groups

First, statistical power was low due to the small group sample sizes; each group had between 13 and 21 participants. In addition, nine dependent variables were used, so the alpha level was adjusted using a Bonferroni correction to control for a Type I error. This adjustment made it more difficult to achieve statistical significance. Second, the participants in the comparison group were of a higher proficiency level than those in the two experimental groups. They already had a relatively high level of fluency (mean length of run and phonation time ratio) at Time 1. The comparison group outperformed the teacher and peer group in terms of having more complex utterances (clauses per AS unit and mean length of AS units) at Time 3. Previous researchers' (e.g., de Jong & Perfetti, 2011; Mojavezi, 2014) suggestion that higher proficiency learners benefit more from task repetition explains why the comparison group benefitted from just performing the 3/2/1 task and why it was difficult for the two experimental groups to outperform the comparison group.

Third, as shown in the previous section, pause length and the use of formulaic language were somewhat related. The statistical analyses revealed that the comparison group produced significantly shorter pauses compared to the teacher and peer group. This result might have been due to the fact that the teacher and peer group used the target formulaic language more frequently. Therefore, the teacher and peer group might have paused more in order to retrieve the target form from memory.

Lastly, the pre-task planning and peer-check activity occurred during the treatment phase and the participants did not have a pedagogical intervention on the recording days at Times 1, 2, and 3; therefore, implementing pre-task planning might have produced more significant differences in the participants' immediate performances. Indeed, previous researchers have usually examined the effects of pre-task planning by comparing group performances in cross-sectional designs immediately after the participants have completed the pre-task planning (e.g., Foster & Skehan,1999; Geng & Ferguson, 2013; Kawauchi, 2005; Mochizuki & Ortega, 2008). In such cases, it is unsurprising that the participants' task performances were influenced by the pre-task planning they had received because learners can connect what they have read in the model passage and retrieve ideas and linguistic forms while planning.

Pre-task planning and L2 learners' performances are strongly related to Levelt's speaking model because planning and repetition occur in the Conceptualizer and Formulator stages, respectively (Skehan, 2015; 2018). When L2 learners have planning time, they can think of ideas in the conceptualizer, which allows them to send the pre-verbal message to the formulator stage, in which they access the mental lexicon for appropriate lemmas that become the basis for syntax building. Skehan (2018) has argued that formulation is a vulnerable stage in L2 speech production because it is underpinned by limited mental lexicon resources and non-automatic syntax-building process (Skehan, 2018, p. 58). These two issues are important because skilled communication requires clear thinking that is then translated into the effective use of the linguistic elements used to express the ideas (Skehan, 2018, p. 114). During the pedagogical intervention stage, the participants engaged in teacher-led pre-task planning, which might have helped them access lexical resources more easily. However, the participants did not have an opportunity for the teacher-led planning time on the recording days. Therefore, the pedagogical intervention did not produce significant differences between groups in terms of the CALF indices.

The Frequency and Types of Target Formulaic Language

Research question 3 concerned the frequency and variety of use of the target formulaic language for the students in the two treatment groups. The average use of the target formulaic language was calculated by dividing the total number of occurrences by the number of participants because the number of participants in each group differed. In addition to counting the frequency and variety of use of the target formulaic language, transcriptions were used to support the statistical evidence. The transcripts illustrate that the participants were able to elaborate the monologues effectively by using the target formulaic language.

Summary of the Frequency and Types of Target Form

Research question 3 produced three main findings. First, the results showed that both the teacher-led and the teacher and peer groups used the target formulaic language more frequently than the comparison group. The comparison group relied on non-target phrases such as *I think* or *because* to give opinions and reasons. The treatment groups' use of the target formulaic language allowed them to produce more detailed monologues compared to their performance at Time 1.

Second, the participants in the teacher-led and the teacher and peer group elaborated on their ideas by giving more examples at Times 2 and 3 compared to Time 1. The target formulaic language *for example* was used more frequently at Time 3. In addition, giving examples was a prominent difference between the comparison group and the two experimental groups. The teacher-led group used *for example* ten times (*M* = .71) and the teacher and peer group did so 23 times (*M* = 1.10) at Time 3. In contrast, the comparison group only gave examples twice (*M* = .15) at Time 3.

Lastly, the transcriptions also revealed that the teacher and peer group produced a wider variety of the target formulaic language than the teacher-led group. For example, the teacher and peer group was able to organize their reasons by saying *It's mainly because….*, *One reason is…*, and *Another reason is…* In contrast, the teacher-led group used one type, *It's mainly because….* to give reasons. This result indicates that the teacher and peer group used a greater variety of the target formulaic phrases because they had been exposed to them or pressured to use them by the peer check activity.

Interpretation of the Frequency and Types of Target Form

The findings concerning the frequency of the participants' use of the target formulaic language raises three important points. First, both text enhancement and the peer check activity helped the participants in the teacher-led group and the teacher and peer group organize, develop, and elaborate their monologues by using the target formulaic language. The excerpts used to answer research question 3 showed that the participants' monologues became better organized as they gave their opinions first and then supported them with concrete reasons or examples. This point is important considering that the comparison group participants did not provide enough examples in their monologues. Megu, who was in the teacher and peer group, noted that it was much easier to speak when she was told the order in which she should express information. The students were told to state an opinion first, and then follow it with reasons. Indeed, many participants stated that telling a monologue for three minutes was a challenging task due to the difficulty of speaking for that length of time. Kenta, who was in the teacher-led group, talked about his strategy for telling a monologue: "I would say opinion first, if I can say a reason, I would say a reason. If it is a bit difficult to tell a reason, in that case, I would show some examples, then I organize a monologue.（最初に意見をいうじゃないですか。で、そっちの意見で、まあー理由を言えそうなやつは理由で、理由がちょっとむずかしいやつって、そういう時は、例示をしてて、そこからたててこうかなって）" Ogawa (2019) reported that Japanese university students felt that it was more difficult to produce a monologue than to take part in small group or pair work tasks because they are unable to receive assistance from other members (e.g., being asked for reasons and examples) in the monologue task. The participants in this study might have felt that organizing their ideas with the target formulaic language helped them produce longer monologues.

Second, both reading the teacher-led models and having a peer-check helped the participants in the teacher and peer group use the target formulaic language during the monologue task possibly because the repetition of the target formulaic language helped the experimental group participants proceduralize the target formulaic language. The results suggest that the pedagogical intervention encouraged the participants to practice the target phrases repeatedly, which plausibly led to a degree of proceduralization. Given that repetition is a key element when incorporating a focus on form element in a task-based classroom (Ellis, 2016), the 3/2/1 task is potentially valuable, as it allows the participants to practice target forms in meaningful contexts (DeKeyser, 2003; Segalowitz, 2003).

Another advantage of the peer check activity is pushing the participants to use the target phrases. When the participants engaged in the 3/2/1 task, I often observed that the speakers paid attention to the formulaic language on the paper the listeners were using to check their use of the phrases. For example, Sumi stated that that the peer check activity encouraged her to use the target formulaic language. While Sumi was speaking during the 3/2/1 task in the beginning of the semester, she frequently looked at the target formulaic language on the paper her partner used. However, as time went by, she gradually had less need to look at the paper to use the target formulaic language. Megu, who was in the teacher and peer group, said that being checked by a peer allowed her to see what phrases she had not used and she felt that she wanted to try to use each type of the target formulaic language.

In this regard, the peer check acted as peer feedback. The feedback pushed the speakers in the teacher and peer group to use a greater variety of the target forms compared to the teacher-led input enhancement. Sumi commented positively about the peer check: “My listener partner sometimes gave me oral feedback that I should have used the target form there. Then, I pay more attention next time.（あそこでいえば良かったのにってみたいなことをポロっていってくれるときがあって、次から気をつけようかな）” Sumi also added that checking her own partner also helped her understand how to use the target form from a listener’s point of view. One advantage of peer feedback is that it serves dual functions to benefit from feedback providers and receivers’ perspectives (Sato, 2017). The participants’ attempts to use the target formulaic language transferred to subsequent 3/2/1 recordings even when a peer did not check them after having been exposed to explicit peer feedback for 13 weeks.

On the other hand, it was difficult for the participants in the comparison group to transfer what they learned outside of the treatment stage into the target tasks although all participants learned the target formulaic language for group discussion purposes outside of the treatment stage. Indeed, the students in the comparison group used the formulaic expressions frequently in group discussion tasks, which was outside of the treatment phases. It is possible to make a link between the results and transfer appropriate processing: memory retrieval that the L2 learners use during acquisition can be best transferred to a testing situation that has similar characteristics to the original learning task (Morris et al., 1977). In this regard, the relationship between the acquisition of the target formulaic language and the monologue recording situation was more appropriate for the experimental groups. In other words, because the pedagogical intervention was provided during the monologue speaking task, its impact on the learners' subsequent target formulaic language use in similar contexts might be more transfer appropriate than the effects of the same target language use in other types of communication.

Development of Communicative Adequacy

Research Question 4 concerned the extent to which the students who received a form-focus pedagogic intervention developed their communicative adequacy over the 13 weeks. Recently, researchers have acknowledged the importance of examining communicative adequacy (Pallotti, 2009; Révész et al., 2016) because CALF measures do not indicate the extent to which learners achieve communicative goals. In this study, communicative adequacy was defined as the degree to which a learner successfully achieves the task's goals in terms of monologue organization and linguistic competence. Unlike Révész et al., who assessed communicative adequacy separately from linguistic competence, communicative adequacy in this study included organization and syntactic complexity, morphosyntactic accuracy, and oral fluency.

Eleven raters rated the participants' two-minute oral performances in terms of organization, syntactic complexity, morphosyntactic accuracy, and oral fluency at Times 1, 2, and 3 using a 5-point rating scale (See Table 9 for the rubric) to answer research question 4. The raw scores were analyzed with the multi-faceted Rasch model. The Rasch person ability estimates were then examined with a repeated-measures ANOVA in order to determine whether communicative adequacy developed significantly during the semester-long treatment. The independent variable was Time (three levels: Time 1, Time 2, and Time 3), and the dependent variable was the Rasch person ability measures.

Summary of the Communicative Adequacy

The main findings for this research question were that all groups significantly improved communicative adequacy between Time 2 and Time 3 and between Time 1 and Time 3 regardless of the treatment condition. All three groups made large, significant gains from Time 1 to Time 3: comparison group (*d* = 3.88), teacher-led group (*d* = 1.82), and teacher and peer group (*d* = 5.15). Furthermore, all groups made large, statistically significant gains from Time 2 to Time 3. These results suggest that the participants learned how to produce more effective monologues after Time 2.

To investigate this issue further, the raw scores awarded by the human raters for each rating category were analyzed descriptively. The descriptive analysis showed two main points. First, the teacher-led group (gain = 25%) and the teacher and peer group (gain = 24%) showed higher gain scores for organization than the comparison group (gain = 18%) between Time 1 and Time 3; Second, the teacher-led group (gain = 31%) and the teacher and peer group (gain = 28%) made greater fluency gains than the comparison group (gain = 12%) between Time 1 and Time 3.

Interpretation of the Communicative Adequacy

The findings for communicative adequacy provide insights concerning the human raters’ perceptions of the participants’ communicative adequacy. All three groups made large, significant communicative adequacy gains. Using Plonsky and Oswald's (2014) benchmarks for Cohen's *d*, all of the groups had a large effect sizes greater than 1.00. The higher communicative adequacy scores for Time 3 imply that the participants, regardless of the treatment, developed communicative adequacy in one academic semester by participating in 3/2/1 tasks every week.

Further descriptive analyses of the gains in each category provided more insight into the development of communicative adequacy. The two experimental groups made more improvement than the comparison group for organization. This improvement was plausibly due to the use of the target formulaic language, given that the experimental groups used the target formulaic language more frequently than the comparison group. This finding provides a link between the use of the formulaic language and gains for organization. If the participants use the target formulaic language such as *in my opinion*, *it is mainly because*, or *for example*, their monologues sound more coherent and better organized to human raters.

Another noteworthy finding is that the two experimental groups made more improvement in perceived oral fluency than the comparison group. The gain in perceived fluency was plausibly due to the link with the use of the target formulaic language as well. The formulaic language helped the learners to speak more fluently because ready-made chunks of correctly formed phrases can be accessed and produced more quickly than generating sentences word by word using syntactic rules (Boers et al., 2006; Wood, 2009, 2010, 2015). In addition, pause location influences the degree to which speakers sound fluent because pauses at a clause or phrase boundary are more closely linked to higher oral fluency than mid-clause pauses.

The human raters' perceptions of the participants' oral fluency somewhat reflected the development of utterance fluency. In research question 1, the repeated-measures ANOVA showed that the teacher and peer group and the teacher-led group significantly developed oral fluency as assessed by mean length of run, while the comparison group did not. This finding, which shows that the human raters were able to detect the participants' ability to speak more fluently, supports previous studies indicating that human raters are sensitive to aspects of oral fluency such as breakdowns and speed (Doe, 2017).

On the other hand, perceived morphosyntactic accuracy, as assessed by the human raters, did not improve (comparison group = 3%, teacher-led group = 12%, teacher and peer group = 4%). This result, which shows that the pedagogical intervention was not associated with perceived morphosyntactic accuracy, is understandable considering that the pedagogical intervention in this study was not specifically focused on the learners' accurate use of morphosyntax. This finding is similar to the findings reported by Boers et al. (2006), who did not find a positive influence for the use of formulaic sequences on perceived accuracy. Furthermore, given that the analytical measure of accuracy, error-free AS units, did not improve over one semester, it is understandable that perceived accuracy did not improve.

The Relationship Among the CALF Indices and Communicative Adequacy

Research question 5 concerned the relationship between the analytical CALF measures and the human ratings of communicative adequacy. Task-based researchers have recently started paying attention to communicative adequacy, but little is known about the relationship between CALF measures and human ratings of communicative adequacy. This relationship has been addressed in only two studies. Révész et al. (2016) reported that filled pauses was the strongest predictor of communicative adequacy followed by speed fluency. Similarly, Sato (2011) reported that perceived fluency was the second strongest predictor of communicative adequacy. In this study, I investigated the relationship between the person ability estimates of communicative adequacy and the CALF measures by conducting a Pearson correlation analysis.

Summary of the Relationship Among the CALF Measures and Communicative Adequacy

Pearson correlation coefficients were computed among the human ratings of communicative adequacy and the nine CALF measures at Times 1, 2, and 3. The results showed that communicative adequacy correlated significantly with two measures at Times 1, 2, and 3: MLR (r = .38, .45, and .33, respectively) and PTR (r = .44, .61, and .60, respectively). In addition, mean length of pauses had mixed results; it correlated with communicative adequacy at Time 1 (r = .49) and Time 2 (r = .51), but correlated negatively at Time 3 (r = -.63). Other fluency measures such as repairs and mean syllable duration had no significant correlation with communicative adequacy.

Further detailed analyses compared two groups in order to better understand why the participants who made long pauses received high communicative adequacy scores from the human raters: (a) the participants who had long mean length of pauses and high Rasch person ability estimates for communication adequacy at Times 1 and 2, and (b) the participants who had relatively long mean length of pauses and low Rasch person ability estimates for communicative adequacy at Time 3. The descriptive analysis suggested that human raters did not perceive pauses which occurred at Time 1 and 2 as disfluent because the mean pause length was relatively short. Another reason is that mid-pause length was more associated with negative perceptions of communicative adequacy, meaning that if a speaker produces relatively long mid-clause pauses, it can influence human raters' perceptions negatively compared to speakers who produced the same number of relatively short mid-clause pauses.

Interpretation of the Relationship Among the CALF Measures and Communicative Adequacy

The results showed that mean length of run and phonation time ratio consistently emerged as factors that correlated significantly for all topics and at each recording time. This result is reasonable because the longer the participants speak and the more fluently they sound, the more positive the human ratings become. This finding implies that oral fluency largely determines who raters view as effective communicators. In previous research, Sato (2011) reported that the standardized regression coefficients showed that content development was the strongest predictor (β = .42) of communicative adequacy followed by fluency (β = .25). Doe (2017) also reported that perceived fluency, as assessed by human raters, correlated strongly with analytical measures of pauses, articulation rate, and phonation time ratio; these findings imply that the human raters were able to detect utterance fluency accurately.

Mean length of pauses had an inconsistent relationship with communicative adequacy. There was a positive relationship at Times 1 and 2, but mean length of pauses had a negative relationship with communicative adequacy at Time 3; thus, as the speakers produced more pauses, they were seen as more able to achieve the task from the raters' point of view at Times 1 and 2. On the other hand, as the speakers paused more, they were perceived as being less successful at Time 3. Usually, pauses indicate disfluency; therefore, the results at Times 1 and 2 were surprising; this result might have occurred because the mean length of pauses was shorter at Time 1 and 2 compared to Time 3. Table 28 shows that all groups produced longer pauses at Time 3 (Comparison group = 1.05, Teacher led = 2.29, Teacher and peer group = 1.57) than at Time 1 (Comparison group = .75, Teacher led = .71, Teacher and peer group = .59) and Time 2 (Comparison group = .86, Teacher led = .72, Teacher and peer group = .62). The speakers' relatively short pauses at Times 1 and 2 did not give the raters a negative impression.

Another possible reason for the positive correlation between pause length and communicative adequacy at Times 1 and 2 and the negative correlation at Time 3 is that the location of the pauses gave raters different impressions. This possibility has been confirmed by previous researchers, who have found that it is important to analyze pause location in addition to pause length (De Jong, 2018; Doe, 2017, Kahng, 2014) in part because L2 speakers pause more often within syntactic clauses (Kahng, 2014; Tavakoli, 2011) and within AS-units (De Jong, 2016; Skehan & Foster, 2005) than native English speakers. To investigate this issue further, I looked at the location of the pauses from the transcripts. The number of pauses each speaker produced and instances of pausing in the middle of a clause were counted. For instance, in "when (0.72) I was in kendo match (0.91) I hit the target." The first pause (0.72) was counted as a mid-clause pause, while the second pause (0.91) occurred at a clause boundary.

The finding is somewhat similar to previous studies. Doe (2017) reported that perceived fluency, as assessed by human raters, was highly correlated with pauses, articulation rate, and phonation time ratio. The finding of the current study also suggests that human raters are able to detect different degrees of oral fluency.

Révész et al. (2016) found that breakdown fluency (pauses) significantly impacted human ratings of communicative adequacy. Pauses can give raters the impression that speakers' process the L2 slowly (Sato, 2014; Sato & McNamara, 2018); thus, the speaking performances were rated as less communicatively adequate if the speakers produced longer pauses. In addition to previous researchers' findings, the follow-up analysis in this study indicated that the frequency of pausing does not always imply disfluency; rather, pause length within a clause or within an AS-unit resulted in lower scores by the human raters. Thus, pauses *between* clauses are indicators of the conceptualization and content planning, whereas pauses *within* clauses signal breakdowns in lexical and syntactic encoding in the formulator (Lambert et al., 2017). In this regard, pauses should be carefully examined depending on (a) the overall length of the pauses and (b) the length of mid-clause pauses. Sato and McNamara conducted retrospective interviews with their human raters and reported that pauses in the middle of the utterance made the speech difficult to comprehend and distracted the raters from the comprehension of the content. The findings of this study support the previous findings that perceptions of oral fluency influence listeners' judgements of comprehensibility (Saito, Trofimovich & Isaacs, 2015; Sato, 2011; Sato & McNamara, 2018).

In contrast to the presence of mid-clause pauses, the speakers' ability to achieve task goals effectively was not associated with the use of more syntactically complex utterances. Human ratings of communicative adequacy had weak but statistically significant correlations with morphosyntactic accuracy at Time 1 and with lexical diversity at Time 2, but the results were not consistent throughout the three testing times. In an opinion-based task, conveying meaning clearly might be considered more important than producing syntactically complex and morphosyntactically accurate sentences. In other words, a lack of complexity might not capture raters' attention in opinion-based monologues. Native-like morphosyntactic accuracy and syntactic complexity were not crucial for communicative success from the raters' point of view (Sato & McNamara, 2018). The findings of the current study suggest that successful oral communication might depend more on comprehensibility rather than on linguistic errors or a lack of syntactic complexity.

Learners' Prioritization While Performing the Monologue Speaking Task

Research question 6 concerned what the learners prioritized during their monologues. This research question was motivated by a desire to fill in the following research gaps. Few researchers have analyzed students' speaking performances qualitatively because TBLT researchers have mainly analyzed students' oral performance quantitatively based on CALF measures (e.g., Boers, 2014; De Jong et al., 2012; Geng & Ferguson, 2013; Ortega, 1999). In addition, few TBLT researchers have examined what learners prioritize when performing speaking tasks.

This research question was answered by administering a retrospective questionnaire. The participants in each group indicated what they prioritized during the monologue recording immediately after completing each recording by choosing one or multiple answers from the following five options: content, grammar, vocabulary, organization, and formulaic language. They then wrote descriptions in Japanese describing what they did.

Summary of the Learners' Prioritization While Performing the Monologue Speaking Task

The descriptive analysis of the frequency data indicating what the speakers prioritized during the 3/2/1 recording revealed that most participants prioritized content. When the speakers prioritized the content of their monologues, they thought about stories related to the topic or related to their own experience. Two speakers thought about the listener's perspective because they wanted the monologue to make sense to the listener. There were 49 instances of prioritizing organization such as speaking logically, elaborating on an idea, and managing time effectively to organize their monologue. The speakers thought about grammar or lexical items 48 times; however, 13 speakers thought about using simple grammar or high-frequency lexical items. There were 73 instances of participants thinking about the target formulaic language.

Interpretation of the Learners' Prioritization While Performing the Monologue Speaking Task

The retrospective questionnaire findings showed that the participants focused on meaning more than form while engaging in the 3/2/1 monologue task. That result is similar to previous findings. For example, the learners in Sangarun's (2005) study were more likely to focus on meaning rather than form during pre-task planning regardless of the planning type. Ogawa (2019) also found that low-proficiency students focused on meaning during the 3/2/1 training.

First, Levelt's speech model (1989) provides an explanation for learners' tendency to prioritize meaning over form. According to Levelt, speakers think of what to say in the conceptualizer stage. Without first thinking of what to say, speakers will be unable to express their ideas using linguistic forms accessed in the formulator and articulation stages. The strategy of attending to content is understandable given that most people primarily focus on meaning in communicative situations. The participants answered that they were concerned about whether they were able to convey their meanings to the listener successfully. Ortega (2005) stated that the presence of authentic listeners helped her participants to speak more comprehensibly in order to meet the listeners' needs. Unlike the regular 3/2/1 training, the participants in this study recorded their speaking individually. Even though the participants did not have an authentic listener when making recordings at Times 1, 2 and 3, they might have thought that I would listen to the recording later; thus, they wanted to prioritize the message first.

Second, the participants in this study were low-proficiency English speakers. Most Japanese university students have little experience speaking English extensively prior to entering a university; therefore, it is challenging for them to pay attention to both meaning and form. For high-proficiency speakers, formulation might be largely automatic and allow for parallel processing, while for lower-proficiency speakers, the formulator stage (lexical retrieval and grammatical encoding) requires attentional resources, and this results in degrees of breakdown in parallel processing (De Bot, 1992; Kormos, 2006). Skehan (1996, 1998) also argued that learners are typically preoccupied with thinking about task content and they therefore do not necessarily pay attention to morphosyntactic accuracy. Ogawa (2019) reported that three low-proficiency Japanese university students did not pay attention to linguistic form during the 3/2/1 training session. Even though they knew that they would not be able to speak grammatically correctly, they prioritized conveying their message to the listeners. Low-proficiency students cannot attend to meaning and form simultaneously (Anderson, 1995; Sangarun, 2005; Yuan & Ellis, 2003), a situation that can lead them to prioritize one aspect of language, typically meaning, to achieve a communicative goal.

To achieve a communicative goal, the participants focused on their own capability to convey meaning; they prioritized what they could say in English rather than what they wanted to say. Twenty-one participants stated that they chose to talk about ideas they could express easily. One strategy they used to prioritize what they could say was relating the topic to their own experiences. It was relatively easy for them to talk about high school experiences and club activities. A second strategy the participants used to maximize their capability was to use simple grammatical constructions and/or high-frequency vocabulary. Thirteen students answered that they used easy syntactic forms and avoided the use of complex grammar. Kenta, who was in the teacher-led group, said, “I feel that I am more able to develop my skills that I can put my opinion in English faster. If I use difficult opinion, I cannot use difficult words. So, I would rather use easy English words and tell my listener partner.（自分の意見をいかに早く翻訳できるか、かなり強まったかなって。むずかしい意見を言っても、やっぱり、むずかしい単語って使えないじゃないですか。だから、なるべく簡単な単語を並べて、いかに相手に伝えられるかっていうところをわりと考えるようにはなったかな）”

Having said that, some participants focused on form while speaking. According to Levelt's model, speakers need to access linguistic knowledge via lexical retrieval and grammatical encoding in the formulator stage. This type of form/meaning mapping can be challenging for many L2 learners. Indeed, 48 of the participants who paid attention to linguistic form commented that they felt that they were unable to express themselves grammatically accurately even when they attended to linguistic form; thus, simply paying attention to lexical or grammatical retrieval does not always lead success. Megu, who was in the teacher and peer group, said that she thought about the content of her last recording, but when it comes to speaking, she was unable to articulate her thoughts well by retrieving appropriate vocabulary. She said, "I cannot come up with the vocabulary.（単語が出てこない）" Sumi also said that she had something to say but she could not put it in English well, therefore, she stopped. It is possible that because the linguistic encoding processes of lower-proficiency learners are less automatized (Lambert, et al, 2017), Megu and Sumi felt frustrated when they were unable to allocate attentional resources effectively for conceptual and linguistic processing.

Sangarun (2005) explained that the participants who focused on both meaning and form during pre-task planning successfully decreased working-memory load, and this allowed them to place more attention on morphosyntactic accuracy during the task performance. On the other hand, the participants in this study were given 1 minute for planning time prior to the recording, which might have been an insufficient amount of time for thinking about both form and meaning. Therefore, the participants spent more time on content than on linguistic form, which led them to struggle to process form-meaning mappings when recording.

The strategy of achieving a communicative goal by focusing on meaning over form might explain why the participants were more able to develop speaking fluency (e.g., mean length of run, phonation time ratio) than morphosyntactic accuracy and lexical diversity. Sumi stated that she paid some attention to grammar, but she thought that trying not to stop speaking was more important. She said, “I try to think to talk something, talk something.（なんかしゃべろう、しゃべろうっていうほうが気にしているかんじ。）” The participants’ prioritization of meaning might be associated with their speaking performances.

Pedagogical Implications

Task-based language classrooms are goal-oriented, meaning-focused, and student-centered; these characteristics often raise questions about how and where learners can learn and practice new target linguistic forms. There is still a gap between TBLT research and classroom teachers’ needs given that many previous task-based studies have been conducted in laboratory settings or outside authentic classroom contexts (Samuda, 2015; Van den Branden, 2016). Because some teachers use TBLT in the classroom, teachers should receive more attention in task-based research (Van den Branden, 2016). In addition, many TBLT studies have been conducted in ESL contexts (e.g., De Jong & Perfetti, 2011; Foster & Skehan, 1999; Geng & Ferguson, 2013). This study was conducted in intact classes in a Japanese university; therefore, the findings provide pedagogical implications that are applicable to language classrooms in EFL contexts.

First, a form-focused intervention using formulaic language might improve L2 speakers’ oral fluency in one academic semester. The 3/2/1 task was already known as an effective way to improve oral fluency because lexical overlap during repetition in a shrinking time condition can enhance fluency development (De Jong & Perfetti, 2011; Nation, 1989; Thai & Boers, 2016). As previous researchers have found the positive impact of teaching formulaic language on L2 oral fluency development (e.g., Tavakoli & Hunter, 2018; Wood, 2009, 2015; Wray, 2002, 2013), this finding also suggests that automaticity is best achieved by the repeated use of target linguistic forms in authentic communicative contexts (DeKeyser, 2003; De Ridder et al., 2007; Segalowitz, 2003). In addition, researchers have investigated learners' oral fluency development in ESL contexts (e.g., De Jong & Perfetti, 2011) or study abroad contexts (e.g., Freed, Segalowitz, & Dewey, 2004; Tavakoli et al., 2016). This study suggests that L2 speakers can develop oral fluency in an EFL context if they are given appropriate instruction and opportunities to practice the target formulaic language in meaningful contexts.

The findings also suggest that providing model input helps students understand how to produce higher quality monologues and use formulaic expressions more effectively; these improvements highlight the effective impact of instruction on the development of oral fluency. Taking notice of a given instance of formulaic language once or twice is not enough to leave durable memory traces (Boers et al., 2006). This means that L2 learners can uptake formulaic language through input flooding or ensuring that the same sequence recurs several times in a relatively short stretch of discourse (Boers et al., 2006). In the beginning of the first semester at university, the participants could not continue speaking for three minutes in English. The teacher's model passage functions as an awareness-raising model for students. In addition, text enhancement of the target formulaic language plays an important role in focusing students' attention on target linguistic form and possibly on form-meaning-function mappings.

Second, a peer check can help students pay attention to target formulaic language, which also leads to improvements in oral fluency. The participants in the teacher and peer group used formulaic language frequently, they used a wide variety of target formulaic language, and they produced fluent language, as demonstrated by longer runs and higher phonation time ratios. Text enhancement during the teacher-led pre-task planning time encourages students to the formulaic language and helps learners speak more fluently. However, awareness-raising itself does not guarantee that learners will use the target phrases in the practice stage. On the other hand, peer monitoring can push speakers to raise their awareness of linguistic form because it can function as oral feedback in which speakers can clearly understand what forms they have (not) used. The explicitness of the peer-check pushed the participants in the teacher and peer group to use a wider variety of target formulaic languages than the members in the other two groups. L2 speakers tend to have a small repertoire of expression and lexis. Without pressure, they might overuse familiar expressions, which might sound monotonous.

Peer feedback can be beneficial from a listener's point of view as well. As Sato (2017) recognized of the dual functions of the peer feedback, L2 learners can learn by monitoring their partners. Indeed, the retrospective interview findings indicated that some of the participants reported that they learned how to use the target formulaic language by checking their speaker partners. Usually, peer corrective feedback on grammatical errors might be challenging for some learners due to their lack of confidence in their linguistic knowledge and group dynamics (Philp et al., 2010). While corrective feedback requires the students to be trained enough to provide the grammatical errors effectively, this peer check might be easier to be implemented regardless of the participants' proficiency level because they can pay attention to whether the target forms are used or not. Rather than pointing out their partner's grammatical errors, peer feedback in this study might be easier for L2 learners. Therefore, from a teacher's point of view, this type of peer feedback might be convenient to implement and it might create a positive mindset in some students as they learn to use new language.

Another noteworthy finding is that these types of form-focused pedagogical interventions can be added to the 3/2/1 task without negative effects on CALF. This study was one of the first studies in which the effects of form-focused instruction during the 3/2/1 task was examined; the participants were not overwhelmed by the additional form-focused attention in the 3/2/1 task. This finding suggests that the 3/2/1 task can be beneficial enough by itself, but repeated practice of formulaic language can lead to improvements in the mean length of run and phonation time ratio without harming the learners' other aspects of CALF.

Lastly, teachers can include instruction of how to achieve communicative goals effectively in opinion-based monologue tasks. One way to help students achieve communicative goals is to teach them discourse organization skills. Fluency and organization were more important criteria for achieving higher communicative adequacy than morphosyntactic accuracy and syntactic complexity from the raters' point of view; therefore, students need to work on these two aspects rather than focus on speaking with a high degree of grammatical accuracy. Many students find it challenging to continue speaking in a monologue because they do not have sufficient linguistic knowledge to express their ideas and they have not proceduralized some of their linguistic knowledge. If students understand how to organize their opinions effectively by using formulaic language, they might be able to sound more proficient and communicatively adequate.

Monologue speaking tests can be conducted in many different ways such as picture description, retelling, or narration, but when they are opinion-based monologues, it is crucial to include organization as one criterion. Indeed, some English proficiency tests (e.g., TEAP, EIKEN, TOEFL iBT) include opinion-based monologue tasks in which test-takers must talk about a preference or opinion about a certain topic. Because more school administrators in Japan have announced plans to include an oral performance test in their entrance exams (e.g., high schools under the Tokyo Metropolitan Board of Education or the National Center for University Entrance Examination), it is important for younger Japanese learners such as junior high school and high school students to prepare for such speaking tests. To do that, classroom teachers should include speaking assessments in the English-language curriculum.

Although task-based language instruction is student-centered, it does not mean that the teacher's role is unimportant; rather, teachers should serve as language guide facilitators (Willis, 1996). Teachers play a vital role in planning and designing lessons and they are crucial when it comes to providing target language input and form-focused instruction, both of which provide learners with opportunities to develop their language skills (Van den Branden, 2016). The results of this study indicate that L2 learners might be able to develop their speaking performances over one academic semester in an EFL context by combining form-focused instruction with the 3/2/1 task.

Summary

In this chapter, I have summarized the results for each of the six research questions in the study. I have also interpreted the findings with reference to the results from previous studies and with follow-up analyses using interview data. I also discussed pedagogical interventions that classroom teachers might utilize to help students improve their speaking proficiency. In Chapter 6, I briefly summarize the main findings of the study. I then discuss the limitations of the study, make suggestions for future research concerning speaking proficiency development, and provide final comments.

CHAPTER 6

CONCLUSION

In this chapter, I briefly summarize the main findings of the study. I then discuss the limitations of the study and make suggestions for future research in second language speaking development studies. Finally, I provide conclusions about the role of teaching formulaic language in EFL settings.

Summary of the Findings

This study was conducted to investigate the effectiveness of form-focused instruction on the longitudinal development of Japanese university students' speaking performances. This study produced six main results. First, the two experimental groups developed mean length of run significantly, the teacher and peer group improved phonation time ratio significantly, and the two experimental groups developed greater oral fluency, possibly because of the instruction of the target formulaic language. However, in contrast to the students in Vercellotti's (2017) longitudinal study, the participants did not improve on all of the CALF indices.

The second finding was that the comparison groups outperformed the experimental groups in the following ways. First, the comparison group participants were able to produce longer runs and sound more fluent than the teacher and peer group at Time 1. Second, the comparison group was able to produce more syntactically complex utterances compared to the two experimental groups. Third, the comparison group was able to produce significantly fewer pauses than the teacher and peer group. and fewer pauses compared to the teacher and peer group at Time 3. These findings indicate that the teacher and peer group initially had lower speaking proficiency and the comparison group participants were relatively fluent before starting the treatment. This starting difference made the findings for research question 1 and 2 difficult to interpret in that the results might have differed if the three groups had started at the same oral proficiency level.

The third finding is that that both the teacher-led and the teacher and peer groups used the target formulaic language more frequently than the comparison group. Furthermore, the teacher and peer group produced a wider variety of the target formulaic language than the teacher-led group. For example, the teacher and peer group used target forms such as *It's mainly because….*, *One reason is…*, and *Another reason is…*, while the teacher-led group used one type, *It's mainly because….*, to give reasons. This result suggests that the peer-check played an important role concerning the use of target forms, while text enhancement only increased noticing of the forms; therefore, the participants in the teacher and peer group had more pressure to use a wider variety of formulaic language, and this pressure led to higher quality performances.

The fourth finding was that all groups significantly improved communicative adequacy between Time 1 and Time 3, and Time 2 and Time 3. The further descriptive analyses of the gains for each criterion showed that the two experimental groups made more improvement than the comparison group for organization scores and fluency scores. The pedagogical intervention positively impacted the participants in the two experimental groups in terms of better organization and oral fluency development as perceived by the human raters.

The fifth finding showed that communicative adequacy correlated strongly with two oral fluency measures, mean length of run and phonation time ratio. In addition, pause length within clauses and AS-units had a negative impact on the human raters' impressions. This finding suggest that fluency plays an important role in giving a positive impression to human raters. This study supports previous studies showing that disfluency markers such as long pauses within clauses make speech more difficult to comprehend and can distract raters from the comprehension of the content (Sato & McNamara, 2018).

The final finding was that that the participants often focused on meaning rather than form during their performances. This finding supports Levelt's (1989) speech model. The participants used the strategy of achieving their communicative goal primarily by focusing on meaning and thinking of what they were able to express in English. This result suggests that it might be difficult for low-proficiency Japanese students to attend to both form and meaning simultaneously.

In sum, the findings supported previous research indicating that formulaic language instruction helps L2 learners' oral fluency development (Tavakoli & Hunter, 2018; Wood, 2009, 2010, 2015; Wray, 2002). By repeating the target formulaic language during the 3/2/1 task, the participants in the two experimental groups might have proceduralized and possibly automatized their speeches; therefore, their mean length of run increased significantly. In addition, teaching formulaic language also helped the learners develop their monologue organization skills; the participants made large gains in organization scores as evaluated by human raters. Moreover, the participants in the teacher and peer group were able to use a wider variety of the target formulaic language compared to those in the teacher-led group. The findings shed light on implementing formulaic language in conjunction with task repetition.

Limitations

This study had four limitations that could have affected the results; therefore, the results should be interpreted with care. The first limitation was that the group *n*-sizes were small; this reduced the statistical power of the quantitative analyses. Second, the participants were in intact classes; therefore, the placement of the participants into groups was not randomized. The third limitation was that the groups differed in terms of English proficiency. The classes were formed based on the students' placement TOEIC test results. I attempted to form three groups whose TOEIC scores did not differ appreciably; however, it was not possible to make all three groups equivalent in terms of English speaking proficiency. The fourth limitation was that some of the measures used in this study (e.g., error-free AS-units) were not sensitive enough to show the participants' linguistic development.

Suggestions for Future Research

Based on the findings produced by this study, several suggestions for future research can be made. The first suggestion is that this study should be replicated with a larger sample size. In the current study, the three groups were made up of 13-21 participants each. The findings would be more robust with a larger sample size of at least 30 participants for each group. Considering that analyzing speaking data is time consuming because of the need to conduct acoustic analyses, few researchers have employed large sample sizes. One way to address this issue is to effectively use the PRAAT speech analysis software (Boersma & Weenink, 2009). In addition to measuring pause length, the software can automatically detect syllable nuclei and estimate the speech rate of performances with a specific script in the program (de Jong & Wempe, 2009). However, in order to allow PRAAT to measure voices automatically, the recorded data should not include background noise. In the current study, the participants' monologues were recorded simultaneously in the same classroom, so other students' performances produced background noise. Ideally, recording should be done in a CALL classroom using individual microphones.

Another suggestion is to combine more types of formulaic language with a different type of task in order to better understand the effects of formulaic language on CALF measures and the perceptions of human raters. Opinion-based monologues were used in the current study, therefore, the target formulaic language concerned stating opinions, providing reasons, and giving examples. One example of other formulaic language types is teaching language for making requests or expressing agreement (Bardovi-Harlig, 2012). High-frequency collocations or lexical bundles can be identified in spoken corpora such as the British National Corpus (BNC) or the Corpus of Contemporary American English (COCA) and taught to students.

The third suggestion is to recruit participants at different English proficiency levels and from different educational settings. The findings of this study indicated that speakers' oral fluency influences human raters' impressions of communicative adequacy, yet more studies need to be conducted to better determine the relationship between L2 learners' command of linguistic features and the degree of communicative success in different tasks, at different proficiency levels, and in different educational settings.

Finally, it would be helpful to use a longer treatment in order to better understand EFL leaners' speaking development. This study is one of the few studies in which a longitudinal design was employed to investigate learners' development in terms of the CALF variables. The current results did not show simultaneous development of all the CALF variables; thus, it might take longer than one semester for Japanese EFL students to improve on all the variables, unlike the ESL learners in Vercellotti's (2017) study. This result is understandable considering that EFL learners have limited opportunities to listen to and produce English both in and outside of classroom. More longitudinal research is needed to identify specific patterns of growth across CALF in an EFL context.

Final Conclusions

As an English learner and English teacher, formulaic language was never the focus in any of my language learning experiences; however, it is becoming increasingly clear that formulaic language is an important element of language learning and use (Conklin & Schmitt, 2012; Wood, 2010; Wray, 2002). Nowadays, English classrooms in Japan are becoming increasingly communicative based on MEXT's guidelines and learners' needs in this era of globalization. As a result, task-based language teaching has been more widely recognized as well. From interlocutors' points of view, successful communication involves more than speaking with accurate grammatical forms; holistic communicative success is necessary when using English as a Lingua Franca communicatively (Sato & McNamara, 2018). In this regard, using formulaic language can help L2 learners feel more confident, sound more fluent, and convey comprehensible messages successfully. I hope that this study leads to more investigations of effective ways to conduct form-focused instruction in a task-based classroom.

REFERENCES

Arczis Web Technologies, Inc. (2019). *Syllable count.* Retrived from http://www.syllablecount.com

Ahmadian, M. J., & Tavakoli, M. (2011). The effects of simultaneous use of careful online planning and task repetition on accuracy, complexity, and fluency in EFL learners' oral production. *Language Teaching Research, 15*(1), 35-59. doi:10.1177/1362168810383329

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

Anderson, J. R. (1987). Skill acquisition: Compilation of weak-method problem solutions. *Psychological Review, 94*(2), 192-210. doi:10.1037/0033-295X.94.2.192

Anderson, J. (1995). *Learning and memory: An integrated approach*. New York, NY: Wiley.

Bardovi-Harlig, K. (2012). Formulas, routines, and conventional expressions in pragmatics research. *Annual Review of Applied Linguistics*, *32*, 206-227. doi:10.1017/S0267190512000086

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

Boers, F., & Lindstromberg, S. (2012). Experimental and intervention studies on formulaic sequences in a second language. *Annual Review of Applied Linguistics, 32*, 83-110. doi:10.1017/S0267190512000050

Boers, F., Eyckmans, J., Kappel, J., Stengers, H., & Demecheleer, M. (2006). Formulaic sequences and perceived oral proficiency: Putting a lexical approach to the test. *Language Teaching Research*, *10*(3), 245-261. doi:10.1191/1362168806lr195oa

Boersma, P., & Weenink, D. (2009). PRAAT: Doing phonetics by computer [Computer Software]. Retrieved from http://www.praat.org

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences*. New York, NY: Psychology Press.

Bosker, H. R., Pinget, A., Quené, H., Sanders, T., & de Jong, N. H. (2013). What makes speech sound fluent? The contributions of pauses, speed and repairs. *Language Testing, 30*(2), 159-175. doi:10.1177/0265532212455394

Brown, J. D. (2014). *Mixed methods research for TESOL*. Edinburgh, Scotland: Edinburgh University Press.

Butler, C. S., (2003). Multi-word sequences and their relevance for recent models of functional grammar. *Functions of Language*, *10*(2), 179-208. doi:10.1075/fol.10.2.03but

Bygate, M., Skehan, P., & Swain, M. (2001). *Researching pedagogic tasks: Second language learning, teaching, and testing*. Harlow, England: Longman.

Conklin, K., & Schmitt, N. (2012). The processing of formulaic language. *Annual Review of Applied Linguistics*, *32*, 45-61. doi:10.1017/S0267190512000074

Cresswell, J. W., & Plano Clark, V. L. (2007). *Designing and conducting mixed methods research*. Thousand Oaks, CA: Sage.

Crossley, S., McNamara, D., & Salsbury, T. (2009). Measuring L2 lexical growth using hypernymic relationships. *Language Learning, 59*(2), 307-334. doi:10.1111/j.1467-9922.2009.00508.x

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

De Bot, K. (1996). The psycholinguistics of the output hypothesis. *Language Learning, 46*(3), 529-555. doi:10.1111/j.1467-1770.1996.tb01246.x

De Jong, N. H. (2016). Predicting pauses in L1 and L2 speech: The effects of utterance boundaries and word frequency. *IRAL*, *54*(2), 113-132. doi:10.1515/iral-2016-9993

De Jong, N. H., & Bosker, H. R. (2013). Choosing a threshold for silent pauses to measure second language fluency. *Proceedings of the 6th Workshop on Disfluency in Spontaneous Speech (DiSS)* (pp. 17-20). Stockholm, Sweden: Royal Institute of Technology.

De Jong, N., & Perfetti, C. A. (2011). Fluency training in the ESL classroom: An experimental study of fluency development and proceduralization. *Language Learning*, *61*(2), 533-568. doi:10.1111/j.1467-9922.2010.00620.x

De Jong, N. H., Steinel, M. P., Florijn, A., Schoonen, R., & Hulstijn, J. H. (2013). Linguistic skills and speaking fluency in a second language. *Applied Psycholinguistics, 34*(5), 893-916. doi:10.1017/S0142716412000069

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

De Jong, N. H., Hulstijn, J. H., Schoonen, R., & Groenhout, R. (2015). Second language fluency: Speaking style or proficiency? Correcting measures of second language fluency for first language behavior. *Applied Psycholinguistics, 36*(2), 223-243. doi:10.1017/S0142716413000210

De Jong, N., & 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*(3), 384-404. doi:10.1177/1362168815606161

De Jong, N. H., & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods*, *41*(2), 385-390. doi:10.3758/BRM.41.2.385

DeKeyser, R. (1998). Beyond focus on form: Cognitive perspectives on learning and practicing second language grammar. In C. Doughty & J. Williams (Eds.), *Focus on form in classroom second language acquisition* (pp. 42-63) New York, NY: Cambridge University Press.

DeKeyser, R. (2001). Automaticity and automatization. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 125-151). New York, NY: Cambridge University Press.

DeKeyser, R. (2003). Implicit and explicit learning. In C. J. Doughty & H. M. Long (Eds.), *The handbook of second language acquisition* (pp. 312-348). Oxford, England: Blackwell.

DeKeyser, R. (2007). Situating the concept of practice. In R. DeKeyser (Ed.), *Practicing in a second language: Perspectives from applied linguistics and cognitive psychology* (pp. 1-18). New York, NY: Cambridge University Press.

De Ridder, I., Vangehuchten, L., & Gómez, M. (2007). Enhancing automaticity through task- based language learning. *Applied Linguistics*, *28*(2), 309-315. doi:10.1093/applin/aml057

Doe, T. (2017). *Oral fluency development activities: A one-semester study of EFL students* (Order No. 10641988). Available from ProQuest Dissertations & Theses database. (2002296277). Retrieved from https://search-proquest-com.libproxy.temple.edu/docview/2002296277?accountid=14270

Doughty, C. (1991). Second language instruction does make a difference: Evidence from an empirical study of SL relativization. *Studies in Second Language Acquisition, 13*(4), 431-469.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly: An International Journal*, *2*(3), 197-221. doi.org/10.1207/s15434311laq0203_2d

Educational Testing Service. (2014). *TOEFL Speaking Rubrics*. Retrieved from https://www.ets.org/s/toefl/pdf/toefl_speaking_rubrics.pdf

Elder, C., & Iwashita, N. (2005). Planning for test performance: What difference does it make? In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 219–238). Amsterdam, The Netherlands: Benjamins.

Ellis, N. C. (1996). Sequencing in SLA: phonological memory, chunking and points of order. *Studies in Second Language Acquisition,1*8, 91-126. doi:10.1017/S0272263100014698

Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition, 24*(2), 143-188. doi:10.1017/S0272263102002024

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

Ellis, R. (2005). Principles of instructed language learning. *System, 33*(2), 209-224. doi:10.1016/j.system.2004.12.006

Ellis, R. (2009a). Task-based language teaching: Sorting out the misunderstandings. *International Journal of Applied Linguistics*, *19*(3), 221-246. doi:10.1111/j.1473-4192.2009.00231.x

Ellis, R. (2009b). The differential effects of three types of task planning on the fluency, complexity, and accuracy in L2 oral production. *Applied Linguistics*, *30*(4), 474-509. doi:10.1093/applin/amp042

Ellis, R. (2010). A framework for investigating oral and written corrective feedback. *Studies in Second Language Acquisition, 32,* 335-349. doi:10.1017/S0272263109990544

Ellis, R. (2016). Focus on form: A critical review. *Language Teaching Research*, *20*(3), 405-428. doi:10.1177/1362168816628627

Ellis, R. (2018). *Reflections on task-based language teaching.* Bristol, PA: Multilingual Matters.

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

Field, A. (2013). *Discovering statistics using SPSS* (4th ed.). London, England: Sage.

Fisher, W. P., Jr. (2007). Rating scale instrument quality criteria. *Rasch Measurement Transactions*, *21*(1), 1095. Retrieved from https://www.rasch.org/rmt/rmt211m.htm

Freed, B. F., Segalowitz, N., & Dewey, D. P. (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*(2), 275-301. doi:10.1017/S0272263104262064

Foster, P., & Skehan, P. (1996). The influence of planning on performance in task-based learning. *Studies in Second Language Acquisition*, *18*(3), 299-324. Retrieved from https:/doi.org./10.1017/S0272263100015047

Foster, P., & Skehan, P. (1999). The influence of source of planning and focus of planning on task-based performance. *Language Teaching Research*, *3*(3), 215-247.

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics*, *21*(3), 354-375. doi:10.1093/applin/21.3.354

Fukuta, J. (2016). Effects of task repetition on learners' attention orientation in L2 oral production. *Language Teaching Research, 20*(3), 321-340. doi:10.1177/1362168815570142

Gass, S., Mackey, A., Alvarez-Torres, M. J., & Fernández-García, M. (1999). The effects of task repetition on linguistic output. *Language Learning, 49*(4), 549-581. doi:10.1111/0023-8333.00102

Gatbonton, E., & Segalowitz, N. (1988). Creative automatization: Principles for promoting fluency within a communicative framework. *TESOL Quarterly, 22*(3), 473-492. doi:10.2307/3587290

Gatbonton, E., & Segalowitz, N. (2005). Rethinking communicative language teaching: A focus on access to fluency. *The Canadian Modern Language Review/La Revue Canadienne Des Langues Vivantes, 61*(3), 325-353. doi:10.1353/cml.2005.0016

Geng, X., & Ferguson, G. (2013). Strategic planning in task-based language teaching: The effects of participatory structure and task type. *System*, *41*(4), 982-993. doi:10.1016/j.system.2013.09.005

Green, S.B., & Salkind, N. J. (2011). *Using SPSS for Windows and Macintosh (6th ed.).* Upper Saddle River, NJ: Pearson Prentice Hall.

Gunnarsson, C. (2012). The development of complexity, accuracy and fluency in the written production of L2 French. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 247-276). Amsterdam, The Netherlands: Benjamins.

Hashemi, M. R., & Babaii, E. (2013). Mixed methods research: Toward new research designs in applied linguistics. *The Modern Language Journal, 97*(4), 828-852. doi:10.1111/j.1540-4781.2013.12049.x

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

Iwashita, N., McNamara, T., & Elder, C. (2001). Can we predict task difficulty in an oral proficiency test? Exploring the potential of an Information—Processing approach to task design. *Language Learning, 51*(3), 401-436. doi:10.1111/0023-8333.00160

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics, 29*(1), 24-49. doi:10.1093/applin/amm017

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher, 33*(7), 14-26. doi:10.3102/0013189X033007014

Johnson, R. B., Onwuegbuzie, A. J., & Turner, L. A. (2007). Toward a definition of mixed methods research. *Journal of Mixed Methods Research, 1*(2), 112-133. doi:10.1177/1558689806298224

Kanda, M. (2015). *Development of English oral proficiency among Japanese high school students* (doctoral dissertation, Temple University Japan). Available from ProQuest Dissertations & Theses database. (1754638039). Retrieved from https://search-proquest-com.libproxy.temple.edu/docview/1754638039?accountid=14270

Kawauchi, C. (2005). The effects of strategic planning on the oral narratives of learners with low and high intermediate proficiency. In R. Ellis (Ed.), *Planning and task-performance in a second language* (pp. 143-166). Amsterdam, The Netherlands: Benjamins.

Kim,Y. (2013). Effects of pretask modeling on attention to form and question development. *TESOL Quarterly*, *47*(1), 8-35. doi:10.1002/tesq.52

Koizumi, R. (2012). Relationships between text length and lexical diversity measures: Can we use short texts of less than 100 tokens. *Vocabulary Learning and Instruction*, *1*(1), 60-69. doi:10.7820/vli.v01.1.koizumi

Kormos, J. (2006). *Speech production and second language acquisition*. Mahwah, N.J: Erlbaum.

Kormos, J., & Trebits, A. (2012). The role of task complexity, modality, and aptitude in narrative task performance. *Language Learning, 62*(2), 439-472. doi:10.1111/j.1467-9922.2012.00695.x

Krashen, S. D. (1994). Self-correction and the monitor: Percent of errors corrected of those attempted vs percent corrected of all errors made. *System*, *22*(1), 59-62. doi:10.1016/0346-251X(94)90040-X

Lambert, C., Kormos, J., & Minn, D. (2017). Task repetition and second language speech processing. *Studies in Second Language Acquisition*, *39*(1), 167-196. doi:10.1017/S0272263116000085

Larsen-Freeman, D. (2006). The emergence of complexity, fluency, and accuracy in the oral and written production of five Chinese learners of English. *Applied Linguistics*, *27*(4), 590-619. doi:10.1093/applin/aml029

Larsen-Freeman, D. (2013). Transfer of learning transformed. *Language Learning*, *63*, 107-129. doi:10.1111/j.1467-9922.2012.00740.x

Lapkin, S., D. Hart & M. Swain (1991). Early and middle French immersion programs–French-language outcomes. *Canadian Modern Language Review–Revue Canadienne Des Langues Vivantes 48*, 11-40.

Laufer, B., & Nation, P. (1999). A vocabulary-size test of controlled productive ability. *Language Testing, 16*(1), 36-55. doi:10.1191/026553299672614616

Lee, S., & Huang, H. (2008). Visual input enhancement and grammar learning: A meta-analytic review. *Studies in Second Language Acquisition, 30*(3), 307-331. doi:10.1017/S0272263108080479

Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning, 40*(3), 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.

Li, L., Chen, J., & Sun, L. (2015). The effects of different lengths of pretask planning time on L2 learners' oral test performance. *TESOL Quarterly*, *49*(1), 38-66. doi:10.1002/tesq.159

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

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement*, *3*(1), 85-106.

Linacre, J. M. (2013). Facets computer program for many-facet Rasch measurement, version 3.71.4. Beaverton, Oregon: Winsteps.com

Long, M. (1991). Focus on form: A design feature in language teaching methodology. In K. De Bot, D. Coste, R. Ginsberg, & C. Kramsch, (Eds.), *Foreign-language research in cross-cultural perspective* (pp. 39-52). Amsterdam, The Netherlands: Benjamins.

Long, M. H. (2016). In defense of tasks and TBLT: Nonissues and real issues. *Annual Review of Applied Linguistics*, *36*, 5-33. doi:10.1017/S0267190515000057

Long, M. H., & Crookes, G. (1992). Three approaches to task-based syllabus design. *TESOL Quarterly*, *26*(1), 27-56. doi:10.2307/3587368

Lyster, R., & Saito, K. (2010). Oral feedback in classroom SLA: Meta-analysis. *Studies in Second Language Acquisition*, *32*, 265-302. doi:10.1017/S0272263109990520

Maad, M. R. B. (2010). Holistic and analytic processing modes in non-native learners' performance of narrative tasks. *System, 38*(4), 591-602. doi:10.1016/j.system.2010.09.013

Mackey, A., & Goo, J. (2007). Interaction research in SLA: A meta-analysis and research synthesis. In A. Mackey (Ed.), *Conversational interaction in second language acquisition: A collection of empirical studies* (pp. 407-452). Oxford, England: Oxford University Press.

Malvern, D., & Richards, B. (2002). Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. *Language Testing, 19*(1), 85-104. doi:10.1191/0265532202lt221oa

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

McNamara, T. F. (1990). Item response theory and the validation of an ESP test for health professionals. *Language Testing*, *7*(1), 52-76. doi:10.1177/026553229000700105

McNamara, T., Knoch, U., & Fan, J. (2019). *Fairness, justice & language assessment.* Oxford: England: Oxford University Press

Ministry of Education, Culture, Sports, Science, and Technology, Japan. (2003). *The action plan to cultivate Japanese with English abilities*. Retrieved from http://www.mext.go.jp/b_menu/shingi/chukyo/chukyo3/015/siryo/04042301/011/002.htm

Ministry of Education, Culture, Sports, Science, and Technology, Japan. (2009). *Shogakko gaikokugo katsudo kenshu gaido bukku* [*Guide book on in-service teacher training for the foreign language activities*]. Tokyo, Japan: Obunsha.

Ministry of Education, Culture, Sports, Science, and Technology, Japan. (2014). *Report on the future improvement and enhancement of English education (Outline): Five recommendations on the English education reform plan responding to the rapid globalization*. Retrieved from http://www.mext.go.jp/en/news/topics/detail/1372625.htm

Ministry of Education, Culture, Sports, Science, and Technology, Japan. (2018). *Gakkou kihon chosa* [*School Basic Examination*] Retrieved Marcy 17, 2019 from http://www.mext.go.jp/b_menu/toukei/chousa01/kihon/kekka/k_detail/1407849.htm

Mochizuki, N., & Ortega, L. (2008). Balancing communication and grammar in beginning-level foreign language classrooms: A study of guided planning and relativization. *Language Teaching Research*, *12*(1), 11-37. doi:10.1177/1362168807084492

Mojavezi, A. (2014). The relationship between task repetition and language proficiency. *Applied Research on English Language. 3*. 29-40. Retrieved from http://are.ui.ac.ir/article_15484_32b2eb98ed12903f9096c297fb465a02.pdf#page=37

Morris, C., Bransford, J., & Franks, J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning & Verbal Behavior, 16*(5), 519-533. doi:10.1016/S0022-5371(77)80016-9

Muranoi, H. (2007). Output practice in the L2 classroom. In R. M. DeKeyser (Ed.), *Practice in a second language: Perspectives from applied linguistics and cognitive psychology* (pp. 51-84). Cambridge, England: Cambridge University Press.

Nassaji, H. (2016). Research timeline: Form-focused instruction and second language acquisition. *Language Teaching, 49*, 35-62. doi:10.1017/S026144815000403

Nation, I. S. P. (1989). Improving speaking fluency. *System, 17*(3), 377-384. doi:10.1016/0346-251X(89)90010-9

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

Nemoto, T., & Beglar, D. (2014). Developing Likert-scale questionnaires. In N. Sonda & A. Krause (Eds.), *JALT2013 Conference Proceedings* (pp. 1-8). Tokyo, Japan: JALT.

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:10.1177/0265532213514401

Nishino, T., & Watanabe, M. (2008). Communication-oriented policies versus classroom realities in Japan. *TESOL Quarterly, 42*(1), 133-138. doi:10.1002/j.1545-7249.2008.tb00214.x

Nishino, T. (2011). Japanese high school teachers' beliefs and practices regarding communicative language teaching. *JALT journal*, *33*(2), 131-155. Retrieved from http://jalt-publications.org/files/pdf-article/art2_19.pdf

Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CALF in instructed SLA: The case of complexity. *Applied Linguistics*, *30*(4), 555-578. doi:10.1093/applin/amp044

Nunan, D. (2004). *Task-based language teaching*. Cambridge, England: Cambridge University Press.

Ogawa, C. (2016). Examining the effects of Types of pretask planning on oral performances. *JALT Journal, 38*(2), 97-118.

Ogawa, C. (2019). Low-proficiency university students' perceptions of pretask planning and their monologue task performances. *Proceedings of the TBLT in Asia 2018 Conference* (pp. 16-25). Kyoto, Japan: JALT TBLT SIG.

Ortega, L. (1999). planning and focus on form in L2 oral performance. *Studies in Second Language Acquisition, 21*(1), 109-148. doi:10.1017/S0272263199001047

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics*, *24*(4), 492-518. doi:10.1093/applin/24.4.492

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

Ortega, L. (2012). *Task-based language teaching in foreign language contexts: One pragmatist's view.* Plenary delivered at the JASELE Conference, Nagoya, August 4, 2012.

Pallotti, G. (2009). CALF: Defining, refining and differentiating constructs. *Applied Linguistics*, *30*(4), 590-601. doi:10.1093/applin/amp045

Pang, F., & Skehan, P. (2014). Self-reported planning behaviour and second language performance in narrative retelling. In P. Skehan (Ed.), *Processing perspectives on task performance* (pp. 95-127). Amsterdam, The Netherlands: Benjamins.

Park, S. (2010). The influence of pretask instructions and pretask planning on focus on form during Korean EFL task-based interaction. *Language Teaching Research 14*(1), 9-26. doi:10.1177/1362168809346491

Philp, J., Walter, S., & Basturkmen, H. (2010). Peer interaction in the foreign language classroom: What factors foster a focus on form? *Language Awareness*, *19*(4), 261-279. doi:10.1080/09658416.2010.516831

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning, 64*(4), 878-912. doi:10.1111/lang. 12079

Prabhu, N. S. (1987). *Second language pedagogy*. Oxford, England: Oxford University Press.

Qi, Y., & Ding, Y. (2011). Use of formulaic sequences in monologues of Chinese EFL learners. *System, 39*(2), 164-174. doi:10.1016/j.system.2011.02.003

Révész, A., Ekiert, M., & Torgersen, E. (2016). The effects of complexity, accuracy, and fluency on communicative adequacy in oral task performance. *Applied Linguistics, 37*(6), 828-848. doi:10.1093/applin/amu069

Robinson, P. (2007). Task complexity, theory of mind, and intentional reasoning: Effects on L2 speech production, interaction, uptake and perceptions of task difficulty. *IRAL—International Review of Applied Linguistics in Language Teaching*, *45*(3), 193-213. doi:10.1515/IRAL.2007.009

Russell, J., & Spada, N. (2006). The effectiveness of corrective feedback for second language acquisition: A meta-analysis of the research. In J. Norris & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 133-164). Amsterdam, The Netherlands: Benjamins.

Saito, K., Trofimovich, P., & Isaacs, T. (2015). Using listener judgments to investigate linguistic influences on L2 comprehensibility and accentedness: A validation and generalization study. *Applied Linguistics, 38*(4), 439-462. doi:10.1093/applin/amv047

Samuda, V. (2015). Tasks, design, and the architecture of pedagogical spaces. In M. Bygate (Ed.), *Domains and directions in the development of TBLT* (pp. 271-301). Amsterdam, The Netherlands: Benjamins.

Samuda, V., & Bygate, M. (2008). *Tasks in second language learning*. Basingstoke, England: Palgrave Macmillan.

Sangarun, J. (2005). The effects of focusing on meaning and form in strategic planning. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 111-141). Amsterdam, The Netherlands: Benjamins.

Sato, M. (2014). Exploring the construct of interactional oral fluency: Second language acquisition and language testing approaches. *System*, *45*, 79-91. doi:10.1016/j.system.2014.05.004

Sato, R. (2010). Reconsidering the effectiveness and suitability of PPP and TBLT in the Japanese EFL classroom. *JALT Journal*, *32*(2), 189-200. Retrieved from https://jalt-publications.org/files/pdf-article/perspectives.pdf

Sato, T. (2011). The contribution of test-takers' speech content to scores on an English oral proficiency test. *Language Testing*, *29*(2), 223-241. doi:10.1177/0265532211421162

Sato, M. (2017). Oral peer corrective feedback: Multiple theoretical perspectives. In H. Nassaji & E. Kartchava (Eds.), *Corrective feedback in second language teaching and learning* (pp. 19-34). New York, NY: Routledge.

Sato, T., & McNamara, T. (2018). What counts in second language oral communication ability? The perspective of linguistic laypersons. *Applied Linguistics,* Advanced online publication. doi:10.1093/applin/amy032

Schmidt, R. W. (1990). The role of consciousness in second language learning. *Applied Linguistics, 11*(2), 129-158. doi:10.1093/applin/11.2.129

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). *The cognitive bases of second language fluency* (1st ed.). New York, NY: Routledge. doi:10.4324/9780203851357

Sharwood-Smith, M. (1993). Input enhancement in instructed SLA: Theoretical bases. *Studies in Second Language Acquisition, 15*(2), 165-179.

Sippel, L., & Jackson, C. N. (2015). Teacher vs. peer oral corrective feedback in the German language classroom. *Foreign Language Annals*, *48*(4), 688-705. doi:10.1111/flan.12164

Skehan, P. (1996). A framework for the implementation of task-based instruction. *Applied Linguistics*, *17*(1), 38-62. doi:10.1093/applin/17.1.38

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

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics, 30*(4), 510-532. doi:10.1093/applin/amp047.

Skehan, P. (2014). *Processing perspectives on task performance: Task-based language teaching 5.* Amsterdam, The Netherlands: Benjamins.

Skehan, P. (2015). Limited attention capacity and cognition: Two hypotheses regarding second language performance on tasks. In M. Bygate (Ed.), *Domains and directions in the development of TBLT: A decade of plenaries from the international conference* (pp. 123-155). Amsterdam, The Netherlands: Benjamins.

Skehan, P. (2018). *Second language task-based performance: Theory, research, assessment*. New York, NY: Routledge.

Skehan, P., & Foster, P. (1999). The influence of task structure and processing conditions on narrative retellings. *Language Learning, 49*(1), 93-120. doi:10.1111/1467-9922.00071

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, The Netherlands: Benjamins.

Spada, N., Jessop, L., Tomita, Y., Suzuki, W., & Valeo, A. (2014). Isolated and integrated form-focused instruction: Effects on different types of L2 knowledge. *Language Teaching Research*, *18*(4), 453-473. doi:10.1177/1362168813519883

Swan, M. (2005). Legislation by hypothesis: The case of task-based instruction. *Applied Linguistics*, *26*(3), 376-401. doi:10.1093/applin/ami013

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). New York, NY: Allyn and Bacon.

Tahira, M. (2012). Behind MEXT's new course of study guidelines. *The Language Teacher*, *36*(3), 3-8. Retrieved from https://jalt-publications.org/sites/default/files/pdf-article/36.3_art1.pdf

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, The Netherlands: Benjamins.

Tavakoli, P., Campbell, C., & McCormack, J. (2016). Development of speech fluency over a short period of time: Effects of pedagogic intervention. *TESOL Quarterly*, *50*(2), 447-471. doi:10.1002/tesq.244

Tavakoli, P., & Hunter, A. M. (2018). Is fluency being 'neglected' in the classroom? Teacher understanding of fluency and related classroom practices. *Language Teaching Research*, *22*(3), 330-349. doi:10.1177/1362168817708462

TextInspector. (2019). Online lexis analysis tool at http://textinspector.com/workflow

Thai, C., & Boers, F. (2016). Repeating a monologue under increasing time pressure: Effects on fluency, complexity, and accuracy. *TESOL Quarterly*, *50*(2), 369-393. doi:10.1002/tesq.232.

Towell, R., Hawkins, R., & Bazergui, N. (1996). The development of fluency in advanced learners of French. *Applied Linguistics, 17*(1), 84-119. doi:10.1093/applin/17.1.84

Vercellotti, M. L. (2017). The development of complexity, accuracy, and fluency in second language performance: A longitudinal study. *Applied Linguistics*, *38*(1), 90-111. doi:10.1093/applin/amv002

Van den Branden, K., Bygate, M., & Norris, J. M. (Eds.). (2009). *Task-based language teaching: A reader*. Amsterdam, The Netherlands: Benjamins.

Van den Branden, K. (2016). The role of teachers in task-based language education. *Annual Review of Applied Linguistics*, *36*, 164-181. doi.org/10.1017/S0267190515000070

VanPatten, B. (1990). Attending to content and form in the input: An experiment in consciousness. *Studies in Second Language Acquisition, 12*(3)*,* 287-301. Retrieved from https://doi-org/10.1017/S0272263100009177

Wigglesworth, G. (1997). An investigation of planning time and proficiency level on oral test discourse. *Language Testing*, *14*(1), 85-106. doi:10.1177/026553229701400105

Willis, J. (1996). *A framework for task-based learning*. Harlow, England: Longman.

Witton-Davies, G. (2014). *The study of fluency and its development in monologue and dialogue.* (Doctoral dissertation, University of Lancaster). Retrieved from http://www.forex.ntu.edu.tw/en/files/writing/4092_dc0088cd.pdf

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*(1), 39-57. Retrieved from https://search-proquest-com.libproxy.temple.edu/docview/85716551?accountid=14270

Wood, D. (2010). *Formulaic language and second language speech fluency.* London, England: Continuum.

Wood, D. (2015). *Fundamentals of formulaic language: An introduction*. London, England: Bloomsbury.

Wray, A. (2000). Formulaic sequences in second language teaching: principle and practice. *Applied Linguistics, 21*(4), 463-489. doi:10.1093/applin/21.4.463

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge, England: Cambridge University Press.

Wray, A. (2008). *Formulaic language: Pushing the boundaries*. Oxford, England: Oxford University Press.

Wray, A. (2013). Formulaic language. *Language Teaching, 46*(3), 316-334. doi:10.1017/S0261444813000013

Yu, X. (2009). A formal criterion for identifying lexical phrases: Implication from a classroom experiment. *System, 37*(4), 689-699. doi:10.1016/j.system.2009.09.012

Yuan, F., & Ellis, R. (2003). The effects of pre-task planning and on-line planning on fluency, complexity and accuracy in L2 monologic oral production. *Applied Linguistics*, *24*(1), 1-27. doi:10.1093/applin/24.1.1

APPENDIX A
CONSENT FORM (JAPANESE VERSION)

同意書

私、小川知恵は、二言語習得に関するデータを収集しており、皆さんにご協力のお願いをいたします。

今学期、皆さんから同意書を提出して頂いた上でデータの収集を行いたいと考えております。収集したデータは下記の目的のためにのみ使用し、本データに基づく出版・発表が行われる場合においても、皆さんの個人情報については個人が特定されることがないよう機密厳重に取扱管理いたします。

尚、データ収集協力は参加者の自由意志に基づくもので、協力を断ることも可能です。また、同意書提出後、途中で協力を辞退することもできます。データの提供に同意していただける場合は、下の欄に署名をしてください。署名がない場合は同意しなかったものとみなし、データ収集は行いません。

データ収集目的:	3/2/1 speaking activity の事前指導がどのように言語習得に寄与するのか研究するため
データの種類:	音声録音
データの内容:	Speaking の録音 質問紙
データ収集日程:	Week 2，8，13，14

※ご質問がある場合は、同意書提出前にお尋ね下さい。データ収集開始後に質問がある場合は、直接下記メールアドレス宛に質問事項を送付して下さい。（日本語可）

oxxxxx@xxxxx.ac.jp

2016年　4月　1日

- 私は、データ収集の趣旨を理解し、上記データ収集に協力することに同意します。
- 私は、本同意書提出後、いつでも協力を辞退できることを理解しています。
- 私は、参加者・講師署名済みの原本を後日受け取ること、同コピーは英語ディスカッション教育センターが保管していることを理解しています。

参加者署名（日本語可）　日付

_________年______月______日

講師署名　日付

_________年______月______日

APPENDIX B
CONSENT FORM (ENGLISH VERSION)

I, Chie Ogawa, would like to ask for your cooperation regarding SLA data collection.

This semester I would like to collect the data described below from you after your submission of the consent form. The data collected will be used only for the purpose stated below, your personal information will remain confidential and you will have anonymity, even in cases of presentations or publications.

Your participation is completely voluntary. You do not have to be involved in this data collection if you do not want to. Furthermore, even if you have agreed to take part, you can withdraw at any time. The data collection will not take place if you do not sign below. If you agree to provide the data stated above, please sign your name below.

Data Collection Purpose:	To explore how pre-task planning of the 3/2/1 speaking activity contributes to language acquisition.
Data Type:	Audio Recording
Data Content:	Recoding students' speaking performance Questionnaire
Data Collection Schedule:	Week 2, 8, 13, 14

※ If you have any questions, please ask me before signing. If you would like to ask questions after signing the form or data collection has started, please email your questions to my email address below (you can write in Japanese).

Date
oxxxxx@xxxxx.ac.jp

- ● I understand the content written on this consent form and I agree to provide the data stated above for this study.
- ● I understand that providing data is voluntary and I can withdraw from the study at any time, even after signing this consent form.
- ● I understand that I will receive the original form later, and that a copy will be kept in the EDC (English Discussion Class) office.

Participant's Signature Date: Year, Month, Day

To be completed on the Japanese version

Instructor's Signature Date: Year, Month, Day

To be completed on the Japanese version

APPENDIX C
BACKGROUND QUESTIONNAIRE (JAPANESE VERSION)

【１】名前ローマ字：	
【２】年齢	__________ 歳
【３】性別：	□男性　□女性
【４】専攻：	（　）学部　（　）学科
【５】入学方法	立教大学に入学する際、以下のどの試験方式で受験しましたか。 □一般入試 □附属高校 □推薦入試（自由選抜　指定校推薦　アスリート選抜入試） □その他（詳しく：　）
【６】英語学習歴：	中学校入学以前に、英語を学習したことがありますか。 □はい　□いいえ 「はい」の場合： 場所（例：ECC ジュニア） ______________________________ 期間（例：〇年間、〇才～） ______________________________ 頻度（例：週〇回）______________________
【７】英語クラブ歴	学校で英語クラブに所属した経験はありますか。「はい」なら、何年間活動していましたか。 □はい　（　年間）　□いいえ
【８】英語学習年数：	これまでに何年間英語を学習していますか。 約 ________ 年間
【９】英語で話す家族（親戚）、友人	英語でコミュニケーションをとる家族・親戚・友人はいますか。「はい」と答えた場合どのくらいの頻度で話しますか。（例：イギリス人の父親と毎日英語で話す。アメリカに住んでいるいとこと半年に一回英語で話す）

	□はい、います　（例：週○回）頻度 （　　　　　　　　　　　　　　） □いいえ、いません	
【１０】海外経験：	旅行や勉強等で、少なくとも３ヶ月以上英語圏に滞在したことがありますか。 □はい　　　　　□いいえ 「はい」の場合：場所______________________ 目的______ ________________ 期間（○年間、○才～）______________________________	
【１１】英語資格：	英語資格試験のスコアをお持ちの場合は記入してください。 覚えている範囲で、受験または取得年月をお書きください。	
	（例）その他：TOEFL 英検スコア： TOEIC スコア： その他：	[　　高校３年　　2 月　　取得] [　　　年　　　月　　取得] [　　　年　　　月　　取得] [　　　年　　　月　　取得]

APPENDIX D
BACKGROUND QUESTIONNAIRE (ENGLISH VERSION)

【１】name	
【２】age	___________
【３】gender	□M　　□F
【４】major	（　　　　　　　　　）Department　（　　　　　　　） Major
【５】entrance exam	What entrance exam system did you use? □ Entrance exam □ Attached high school □ Recommendation system （Self-recommendation　　School recommendation　　Athletes recommendation） □ Others（e.g. :　　　　　　　　　　　　　　）
【６】 English learning experience	Have you studied English before junior high school? □ yes　　□no If yes, Place（e.g., : ECC Junior） ____ Length（e.g., : ○years、○years old～） ____ Frequency（e.g., : ○times/week） ________________________________
【７】 English club experience	Have you joined an English club in school? □ Yes（　　　　　years）　　　　□ No
【８】 Length of English learning experiences	How long have you been studying English? About _________ years
【９】 Family or friends who you communicate in English	Do you have any family members or friends whom you speak English with? If yes, how often do you talk with them? (e.g., I have English father and I talk with him every day. I have a cousin who lives in the U.S. and I talk with her twice a year) □ Yes（ : ○ time / week）Frequency（　　　　　　　　　） □ No

<table>
<tr><td rowspan="1">【１０】
Oversea
experiences</td><td colspan="2">Have you stayed in English speaking countries more than three months?
□ Yes　　　　　□ No
If「Yes」　Place_______________________ Purpose_______________________
Length（○Years、Age○～） _______________________________</td></tr>
<tr><td rowspan="2">【１１】
Certificate</td><td colspan="2">If you have any scores of English, please write. (If you remember, specify the scores and the year you received)</td></tr>
<tr><td>（e.g.） Others: TOEFL
EIKEN:
TOEIC:
Others:</td><td>[Grade　　Month February Received]
[Grade　　Month　　　Received]
[Grade　　Month　　　Received]
[Grade　　Month　　　Received]</td></tr>
</table>

APPENDIX E
LIST OF FORUMULAIC PHRASES TAUGHT IN THE DISCUSSION COURSE IN SPRING SEMESTER, 2016

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…
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…
Examples	
Asking for Examples	Giving Examples
For example?	For example / For instance,…
For instance?	One example is…
Can you give me an example?	Another example is…
Joining a Discussion	
Joining a Discussion	Asking Others to Join a Discussion
Can I start?	Who would like to start?
Can I say something?	Does anyone want to comment?
Can I ask a question?	Does anyone want to add something?
Possibilities	
Asking about possibilities	Talking about possibilities
If?	If…
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 {me/ name} ?	

APPENDIX F
3/2/1 RECORDING RETROSPECTIVE QUESTIONNAIRE (JAPANESE VERSION)

【Week 2, 8, 14】スピーキングについての質問

Name: Student ID

このアンケートはみなさんの英語学習歴、および現在の英語力についてよりよく理解するためのものです。アンケートは２つのパートから成り立っています。パート１では、スピーキングについて、パート２ではあなた自身について尋ねています。回答は授業運営の参考に、また研究のために使われますが、成績とは一切関係ありませんので、正直にお答えください。よろしくお願い致します。

〈パート１〉

【1】3 分間-2 分間-1 分間のスピーキングはどのように感じましたか。当てはまる数字に〇をつけてください。また理由も書いてください。

難しかった	比較的難しかった	比較的簡単だった	簡単だった
1	2	3	4

理由：

【2】スピーキングを行う時、どのようなことを意識しましたか？あてはまるものの□にチェック(✔)をつけてください。また具体的にどのようなことを考えたのか記述してください。□は５つあります。（複数可）

□ **話す内容を意識した**

《例：何を話そうか、トピックを考えながら話した。例えば、〇〇について経験が有るのでそれについて話すと簡単だと思った。話しやすいトピックや自分の知っている内容について話した。》

あなたの言葉で具体例を下に書いてください）

□ **文法を意識した**

《例：例えば複数形、過去形、-ing 形など文法が正しいか注意しながら話した。中学校で習ったような簡単な文法で話した。入試対策で文法を勉強したので難しい文法を使った。》

（**あなたの言葉で**具体例を下に書いてください）

□ **ファンクションフレーズを意識した（詳しく下に書いてください）**

□ **語彙や単語を意識した**

《例：英語で、この単語はなんていうのか、ということを考えながら話した。特に、〇〇という単語はたくさん使った。トピックに関連して〇〇という動詞、形容詞、名詞を考えながら話した》

（**あなたの言葉で**具体例を下に書いてください）

□ **話の組み立てについて意識した。**
《例：時間配分を気にしながら話した。例えば、一文で終わるのではなく、もっと話を膨らまそうとした。起承転結を考えた。論理的に話ができるように気をつけた。》
（**あなたの言葉で**具体例を下に書いてください）

□ その他（**あなたの言葉で**具体例を下に書いてください）

APPENDIX G
3/2/1 RECORDING RETROSPECTIVE QUESTIONNAIRE
(ENGLISH VERSION)

Name:____________________________ Student ID________________________
This questionnaire is for the purpose of understanding more about your English learning history and your current English ability. The questionnaire consists of two parts. Part 1 asks about speaking, and Part 2 asks about yourself. The responses are used as the reference to class management and also for research, but there is no relation between your response and the course grade; thus, please answer honestly. Thank you for your cooperation.

1. How did you feel about the 3-minute-2-minute-1-minute speaking task? Circle the appropriate number. Also please write the reasons

Difficult	Relatively difficult	Relatively easy	Easy
1	2	3	4

Reason:

2. When you spoke, what did you pay attention to? Please put checkmark ✔in the box □. Also please describe what you thought about specifically. *There are five check boxes*. (Multiple answers are possible)

□ I paid attention to the content of what I spoke about. (Please explain more in detail in your own words below)
(e.g., I spoke while thinking about the topic. I was thinking about the topics. For example, I thought it is easy to talk about XX because I have experienced (the XXX). I talked about topic(s) that were easy to talk about or content I know.)

□ I focused on the grammar. (Please explain more in detail below)
(e.g., For example, I spoke while paying attention to whether grammar, such as plurals, past tense, and *-ing* forms, are accurate. I spoke using easy grammar such as the grammar I learned in junior high school. I used difficult grammar because I studied difficult grammar preparing for the entrance examination.)

□ I focused on functional formulaic phrases. (Please explain more in detail below)

□ I focused on vocabulary.
e.g., I spoke while thinking about what the Japanese word is in English. I used the xx especially frequently. I spoke while thinking of verbs, adjectives, or nouns of xx related to the topic.

□ I focused on the speech organization (Please explain more in detail below)
(e.g., I tried to manage my time. For example, not only finishing with one sentence but I try to expand. I also organize my speech. I try to be more logical.)

□ Other (Please explain in detail below)

APPENDIX H
3/2/1 TRAINING REFLECTION QUESTIONNAIRE (JAPANESE VERSION)

【Lesson 13】3/2/1 task についてのアンケート; Name: ____________________________

授業では３分間、２分間、１分間を話す練習をしました。その 3/2/1 minute の活動について伺います。成績には関係ないので率直な意見をお願いします。

次の各文にどの程度同意しますか。設問について、あてはまる番号を一つ選んで〇をして下さい。また下線部に理由を書いてください。

1. 3/2/1 分を話すアクティビティは得意である。

1	2	3	4	5	6
全くそう思わない	そう思わない	あまりそう思わない	少しそう思う	そう思う	強くそう思う

理由を下に書いてください

2. ３分間話す時、どのようなことに気をつけていますか？できるだけ詳しく教えてください。

3. １分間話す時、どのようなことに気をつけていますか？できるだけ詳しく教えてください。

4. 3/2/1 分のアクティビティに対して、感想を書いてください。できるだけ詳しく率直に書いてください。

プランニングについて質問します。3/2/1 分の前に教師の例文(Model passage)を聞きました。

5. **教師の例文は必要である。**

1	2	3	4	5	6
全くそう思わない	そう思わない	あまりそう思わない	少しそう思う	そう思う	強くそう思う

理由を書いてください

6. **教師の例文を参考にしている。**

1	2	3	4	5	6
全く参考にしない	参考にしない	ほとんど参考にしない	少し参考にする	参考にする	とても参考にしている

理由を書いてください。具体的にどのような点を参考にしますか。

ペアチェックについて質問します。3/2/1 分の間にパートナーから、ファンクションフレーズのチェックを受けました。

7. **ペアチェックは効果的だと思う。**

1	2	3	4	5	6
全くそう思わない	そう思わない	あまりそう思わない	少しそう思う	そう思う	強くそう思う

理由を書いてください

APPENDIX I
3/2/1 TRAINING REFLECTION QUESTIONNAIRE (ENGLISH VERSION)

【Lesson 13】Questionnaire about 3/2/1 task

During the lesson, we practiced speaking for 3 minutes, 2 minutes, and 1 minutes. This questionnaire asks about the 3/2/1/ minute activity. Because your answer has no relation to your grades, please provide your honest opinion.

Please read the following sentence. How much do you agree with the following statements? Please select one and circle it. Also please write the reasons below.

1. I am good at speaking in the 3/2/1 speaking tasks.

1	2	3	4	5	6
I strongly disagree	I disagree	I slightly disagree	I slightly agree	I agree	I strongly agree

Please write reasons why you think so below.

2. When you talk for three minutes, what do you focus on? Please write in detail.

3. When you talk for one minute, what do you focus on? Please write in detail.

4. Write your thoughts and opinions of the 3/2/1 speaking performance. Please write in detail.

Here are the questions about planning. Before the 3/2/1/ minute task, you listened to the teacher's example models and passages.

5. Listening to teacher's model example is necessary.

1	2	3	4	5	6
I strongly disagree	I disagree	I slightly disagree	I slightly agree	I agree	I strongly agree

Please write the reasons

6. I use the teacher's model speech as a reference for 3/2/1 task.

1	2	3	4	5	6
I strongly disagree	I disagree	I slightly disagree	I slightly agree	I agree	I strongly agree

Please write the reason(s). Specifically, what point(s) did you use as a reference?

Here are questions about the pair-check. Your partner checked your use of the phrases while you did the 3/2/1-minute task.

7. I think the peer-check was effective.

1	2	3	4	5	6
I strongly disagree	I disagree	I slightly disagree	I slightly agree	I agree	I strongly agree

Please write reasons.

APPENDIX J
INTERVIEW QUESTIONS

- Please tell your impression of 3/2/1. Why did you choose the number?
（3/2/1 の感想を教えてください。なぜ、アンケートでこの数字を選んだのですか。）

- Is it necessary to have thinking time (planning time) before doing the 3/2/1 task?
（3/2/1 の前に考える時間（プランニングの時間）は必要ですか。）

- You plan before doing the 3/2/1 task. What kind of things do you think about when planning?
（3/2/1 の前に、プランニングを行いますが、いつもどのようなことを考えてプランをしていますか。）

- When did you consider the teacher's example sentences? What did you pay attention to in the sentences?
（教師の例文はどういうときに参考にしますか。どういうところに注目していますか。）

- I would like to ask your opinion about the pair-check. What ability do you think can be improved using the pair-check?
（ペアチェックについて意見をお尋ねします。どのような力が高まると思いますか。）

- How was today's recording? How do you think today's 3/2/1 performance was compared to your first recording? Why do you think so?
（今日の録音はどうでしたか。自分は第一回目の録音時と比べて 3/2/1 のパフォーマンスはどのようになったと思いますか。なぜですか。）

APPENDIX K
SAMPLE OF TEACHER-LED MODEL PASSAGE AND PLANNING (WEEK 2)

- Who are your best friends? Why?

  Personally speaking, I think I don't have many friends in university yet. But, I have a very good friend from high school. Her name is Yuriko. One reason why she is a good friend is because we had the same hobby. She was in the same tennis club in high school. I enjoyed playing and practicing tennis with her. Another reason is she is very kind. For instance, when I was sick and absent from school, she kindly gave me her notebook so I could study.

- Who do you often talk to in your family? Why?

  In my opinion, I like to talk to my sister in my family. It is mainly because she is close to my age. She is two years younger than me. So, we can talk about many similar things such as hobby, favorite fashion, and dreams. When I was an elementary school student, we had a lot of fights because of very small things. But now, we are becoming more grown up, so we are becoming the best friends. For example, last month, we went to Tokyo Disney Land together. We enjoyed riding a splash mountain and other rides. Next year, my sister and I are planning to take a trip to Korea together. If I go to Korea, I would like to eat a lot of spicy food.

Let's plan

Please write what you want to say. Please try to write English words.
You don't need to write a full sentence.

	who?	why?
best friend		
who do you often talk to in your family		

APPENDIX L
PAIR CHECK CARD

Week 2
Check card (Opinion)

	3 minutes	2 minutes	1 minute
In my opinion			
Personally speaking, I think			
I'm not sure, but I think….			

Lesson 3
Check card (Opinion, Reason)

	3 minutes	2 minutes	1 minute
Opinion ● In my opinion ● Personally speaking, I think ● I am not sure, but I think			
One reason is…			
Another reason is….			
It's (mainly/ partly) because…			

Lesson 4
Check card (Opinion, Reason)

	3 minutes	2 minutes	1 minute
Opinion ● In my opinion ● Personally speaking, I think ● I am not sure but I think			
Reason ● It's (mainly/ partly) because ● One reason is ● Another reason is			

Lesson 5 (Opinion, Reason)

	3 minutes	2 minutes	1 minute
Opinion ● In my opinion ● Personally speaking, I think ● I am not sure but I think			
Reason ● It's (mainly/ partly) because ● One reason is ● Another reason is			

Lesson 6

Check card (Opinion, Reason, Example)

	3 minutes	2 minutes	1 minute
Opinion ● In my opinion ● Personally speaking, I think ● I am not sure but I think			
Reason ● It's (mainly/ partly) because ● One reason is ● Another reason is			
For example/ For instance			
One/ Another example is…			

Lesson 7

Check card (Opinion, Reason, Example)

	3 minutes	2 minutes	1 minute
Opinion ● In my opinion ● Personally speaking, I think ● I am not sure but I think			
Reason ● It's (mainly/ partly) because ● One reason is ● Another reason is			
Example ● For example/For instance ● One/ Another example is…			

Lesson 8

Check card (Opinion, Reason, Example)

	3 minutes	2 minutes	1 minute
Opinion ● In my opinion ● Personally speaking, I think ● I am not sure but I think			
Reason ● It's (mainly/ partly) because ● One reason is ● Another reason is			
Example ● For example/For instance ● One/ Another example is…			

Lesson 9
Check card (Opinion, Reason, Example)

	3 minutes	2 minutes	1 minute
Opinion ● In my opinion ● Personally speaking, I think ● I am not sure but I think			
Reason ● It's (mainly/partly) because ● One reason is ● Another reason is			
Example ● For example/For instance ● One/ Another example is…			

Lesson 10
Check card (Opinion, Reason, Example)

	3 minutes	2 minutes	1 minute
Opinion ● In my opinion ● Personally speaking, I think ● I am not sure but I think			
Reason ● It's (mainly/partly) because ● One reason is ● Another reason is			
Example ● For example/For instance ● One/ Another example is…			

Lesson 11
Check card (Opinion, Reason, Example, Possibility)

	3 minutes	2 minutes	1 minute
Opinion ● In my opinion ● Personally speaking, I think ● I am not sure but I think			
Reason ● It's (mainly/partly) because ● One reason is ● Another reason is			
Example ● For example/For instance ● One/ Another example is…			
Possibility ● If			

Lesson 12
Check card (Opinion, Reason, Example, Possibility)

	3 minutes	2 minutes	1 minute
Opinion ● In my opinion ● Personally speaking, I think ● I am not sure but I think			
Reason ● It’s (mainly/partly) because ● One reason is ● Another reason is			
Example ● For example/For instance ● One/ Another example is…			
Possibility ● If			

Lesson 13
Check card (Opinion, Reason, Example, Possibility)

	3 minutes	2 minutes	1 minute
Opinion ● In my opinion ● Personally speaking, I think ● I am not sure but I think			
Reason ● It’s (mainly/partly) because ● One reason is ● Another reason is			
Example ● For example/For instance ● One/ Another example is…			
Possibility ● If			

APPENDIX M
3/2/1 QUESTION CARDS TIME 1

Do you think doing club activities is a good idea for students?
(Have you joined a club activity before? What did you learn from your experiences? Why did you decide your club or circle in R university?)

＜クラブ活動（サークル）をすることは学生にとって良い事ですか＞
（クラブ活動に参加した事はありますか。その経験から何を学びましたか。R大学のサークルを選んだ理由など）

これから言いたい内容を考えます。自分の言いたい内容を英語で箇条書きで考えてみましょう。本番中は紙を見ずに、時間以内でたくさん話しましょう。
(Now, it is your planning time. Please brainstorm English words or what you want to say. During recording, try to talk a lot without looking at your planning paper)

APPENDIX N
3/2/1 QUESTION CARDS TIME 2

Do you think eating in is better than eating out?
(Do you often eat out or eat in? What kind of food do you like? Who do you eat dinner with?)

＜家で食べる方が、外食よりも良いと思いますか。＞
(普段、外食が多いですか、家で食べる方が多いですか。どのような食べ物が好きですか。誰と夕食を食べますか。)

これから言いたい内容を考えます。自分の言いたい内容を英語で箇条書きで考えてみましょう。本番中は紙を見ずに、時間以内でたくさん話しましょう。
(Now, it is your planning time. Please brainstorm English words or what you want to say. While recording, try to talk a lot without looking at your planning paper)

APPENDIX O
3/2/1 QUESTION CARDS TIME 3

Do you think learning English is important for students?
(Do you think studying abroad is a good idea for university students?
What are other good ways to improve your English skills?

＜英語を学習する事は大切ですか＞
(留学をする事は良い考えだと思いますか。英語の力を伸ばすためには他にどのような方法がありますか。)

これから言いたい内容を考えます。自分の言いたい内容を英語で箇条書きで考えてみましょう。本番中は紙を見ずに、時間以内でたくさん話しましょう。
(Now, it is your planning time. Please brainstorm English words or what you want to say. During recording, try to talk a lot without looking at your planning paper)

APPENDIX P
HUMAN RATER TRAINING

Thank you so much for helping me with the ratings.

A student is speaking about one of the following topics for 2 minutes. When the student ID number is 1XX, they talk about topic 1. When the ID number is 2XX, they talk about topic 2. When the student ID number is 3XX, they talk about topic 3. Please evaluate their speech based on the rubric.

*****Sorry! My voice is also recoded after the bell rang. Also, some of the recording situations is not as good as other classrooms. Please try to ignore the background noise.

The students talk about one of the following topics.

Topic	Student ID	Questions
Club activity	101-149	Do you think doing a club activity is a good idea for students? Have you ever joined a club before? What did you learn from your experiences? Why did you choose your club in this university?
Eating	201-249	Do you think eating out is better than eating in? Do you often eat out or eat in? What kind of food do you like? Who do you often eat dinner with?
English learning	301-349	Do you think learning English is important for students? Do you think studying abroad is a good idea for students? What are other good ways to improve your English skills?

Please don’t worry too much about the ratings. (As long as you are consistent about your rating, that is good.)

- **Organization**: Please evaluate to what extent a speech is coherent and well-organized. Students’ might answer one or two questions but please evaluate if the speech was organized when answering them.
- **Complexity**: Please evaluate to what extent a speech utilizes complex grammar (e.g., coordination, many clauses)
- **Accuracy**: Please evaluate to what extent a speech has grammatical error-free utterances.
- **Fluency**: Please evaluate to what extent a speech is smoothly delivered (e.g., without hesitation, fillers, repetition).

【Practice】Please write a number for each criteria (1-5).

Student 1: (100)

	Unsuccessful	Poor	Moderately successful	Successful	Very successful
Organization	1	2	3	4	5
Complexity	1	2	3	4	5
Accuracy	1	2	3	4	5
Fluency	1	2	3	4	5

Student 2: (200)

	Unsuccessful	Poor	Moderately successful	Successful	Very successful
Organization	1	2	3	4	5
Complexity	1	2	3	4	5
Accuracy	1	2	3	4	5
Fluency	1	2	3	4	5

Student 3: (201)

	Unsuccessful	Poor	Moderately successful	Successful	Very successful
Organization	1	2	3	4	5
Complexity	1	2	3	4	5
Accuracy	1	2	3	4	5
Fluency	1	2	3	4	5

Student 4: (300)

	Unsuccessful	Poor	Moderately successful	Successful	Very successful
Organization	1	2	3	4	5
Complexity	1	2	3	4	5
Accuracy	1	2	3	4	5
Fluency	1	2	3	4	5