# THE EFFECTS OF EXPLICIT FORM-FOCUSED INSTRUCTION ON L2 ORAL PROFICIENCY DEVELOPMENT

---

A Dissertation
Submitted
to the Temple University Graduate Board

---

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

---

By
Shzh-chen Nancy Lee
December, 2019

Examining Committee Members

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

## ABSTRACT

This study was an examination of the effects of explicit form-focused instruction on the English development of Japanese university students during a seven-week intervention. Speaking proficiency development is a continuous challenge for most EFL learners who have limited exposure to the target language outside the classroom. Within the communicative language teaching paradigm, task-based language learning (TBLT) has been considered an effective approach for developing students' speaking proficiency. However, while TBLT has been increasingly implemented in English language classrooms, investigations in which explicit form-focused instruction has been integrated into task based speaking tasks are limited.

This longitudinal classroom-based study had five purposes. The first purpose was to examine the development of speaking proficiency in terms of syntactic complexity, syntactic accuracy, and oral fluency. The second purpose was to examine the development of speaking proficiency by comparing learners who received form-focused instruction with those who did not receive form-focused instruction by examining differences in their syntactic complexity, syntactic accuracy, and oral fluency development. The third purpose was to quantitatively and qualitatively examine the week-to-week trajectory of changes in speaking proficiency development. The fourth purpose was to examine learner affective variables. The final purpose was to examine the participants' perceptions concerning the development of speaking proficiency based on their self-assessments.

The participants were 104 first-year students enrolled in a Japanese university. All of the participants narrated a different four-picture cartoon in English once a week for nine weeks. The participants were divided into one comparison group and two intervention groups: form-focused instruction (FFI) and form-focused instruction and peer feedback (FFI + PF). The form-focused instruction intervention included ten minutes of grammar instruction focused on three past tense forms: simple past, past continuous, and past perfect, as well as ten minutes of peer feedback. Between the pretest and posttest, the FFI and FFI + PF participants received seven weeks of instruction before their weekly cartoon narration. Participants in the comparison group did not receive any weekly interventions. The pretest and posttest narration data of all participants were transcribed and analyzed using six CAF measurements: mean length of T-unit, clause/T-unit ratio, percentage of error-free T-units, percentage of accurate past tense usage, speech rate, and self-repair. Moreover, the pretest and posttest narrations recorded by all of the participants were analytically rated by three raters. From the sample of 104 participants, nine participants (three representatives from each group) were chosen for a week-to-week trajectory analysis in which their six CAF performances were qualitatively and quantitatively analyzed. All of the participants completed two questionnaires after taking the pretest and posttest. I developed both questionnaires based on previous literature related to second language speaking proficiency development. The English Speaking Learner Affect Questionnaire was used to examine differences in participants' classroom English speaking anxiety, English speaking self-efficacy, and desire to speak English. The Self-Assessment of Speaking Task Questionnaire was used to examine

changes in the participants' self-assessment of their own performances in terms of syntactic complexity, syntactic accuracy, and oral fluency.

The results indicated that form-focused instruction was effective at improving the participants' global syntactic accuracy and their accurate use of the simple past tense. However, form-focused instruction did not lead to improvements in syntactic complexity or oral fluency. However, explicit form-focused instruction did not lead to detrimental effects on syntactic complexity or oral fluency; thus, form-focused instruction did not appear to stimulate trade-off effects between syntactic complexity, syntactic accuracy, and oral fluency. The analytical ratings provided by the raters indicated that form-focused instruction did not lead to significant improvements in terms of syntactic complexity, syntactic accuracy, or oral fluency.

The results also indicated that the form-focused instruction intervention was effective at quickly improving syntactic accuracy because participants who received both form-focused instruction and peer feedback improved more quickly in the simple past compared to the participants who only received form-focused instruction. However, the participants who improved syntactic accuracy quickly might have experienced trade-offs with oral fluency ability, as improvements in speed fluency were not noted until their syntactic accuracy reached a ceiling effect where their global accuracy and accurate use of simple past tense stopped to improve.

The results indicated that form-focused instruction did not influence classroom English speaking anxiety, as there was no significant difference seen between participants who received form-focused instruction and participants who did not. However, an

excessive amount of form-focused instruction might have negative impacts on speaking anxiety because the participants who received both form-focused instruction and peer feedback showed an increase in their classroom English speaking anxiety. On the other hand, form-focused instruction had no impact on English speaking self-efficacy and it had positive effects on the desire to speak English because the participants who received both form-focused instruction and peer feedback exhibited a significant increase in their desire to speak English compared to the participants who did not receive form-focused instruction and the participants who received only form-focused instruction.

Finally, based on the learners' self-assessment, form-focused instruction had no effects on speaking proficiency development as there was no difference in self-assessed syntactic complexity, syntactic accuracy, and oral fluency between the participants in the three groups. Participants who received form-focused instruction interventions did not assess themselves to having higher syntactic accuracy despite their improvements on the CAF measures.

## ACKNOWLEDGMENTS

First, I would like to express my deepest appreciation to my dissertation supervisor Dr. Paul Leeming for his endless support and patience, and for sharing his expertise in task-based language teaching research with me. His firm yet relaxed supervision motivated me to keep going forward. Dr. Leeming completed a mission-impossible mission by watching over me and giving me a gentle push when I needed it. Without his guidance and support, this dissertation would not have been possible.

Second, I would like to thank Dr. David Beglar whose work demonstrated to me just how crucial it is to pursue research in the area of speaking proficiency development. I would like to especially thank him for his generous support throughout the whole research and writing process despite performing an additional 101 duties. He generously read hundreds of my drafts and always gave kind yet constructive feedback to improve my writing and research. His guidance through the multifaceted field of SLA had a lasting effect on me.

Third, I would also like to thank my dissertation defense committee members for working with me to write this publication. I want to thank Dr. Tomoko Nemoto for her generous input, especially in terms of statistics, and Dr. Timothy Doe for guiding me with his expertise in speaking development research. Without the support of all of the Temple faculty members who provided outstanding lectures and incredible insight, passionately sharing their passion for SLA research, my research would never have reached the heights nor explored the depths of my capabilities. I would also like to thank all my cohort members for being great role models and colleagues.

Fourth, I want to thank two friends, Chie Ogawa, for being the best writing/triathlon buddy and cheering for me all the way, and Jennifer Teeter, for always being there for me when I needed her, and for helping me with proof reading. Without these two powerful ladies and their endless unconditional friendship, I would not have made it this far.

Finally, I want to thank my tiger mum who roared when I needed the energy boost and for always making me a stronger person.

Shzh-chen Nancy Lee

Osaka, Japan

Dec, 2019

## TABLE OF CONTENTS

PAGE

## APPENDICES

## LIST OF TABLES

## LIST OF FIGURES

# CHAPTER 1

# INTRODUCTION

## The Background of the Issue

The emergence of English as a lingua franca has made the ability to speak English a global educational goal. However, in many English-as-a-foreign-language (EFL) contexts, learners struggle to develop speaking proficiency because of their limited exposure to the target language and lack of opportunities to use English outside of the classroom. Speaking is therefore considered by some to be a more difficult skill to develop than reading, listening, and writing (Celce-Murcia & Olshtain, 2000; Gan, 2013; Taguchi, 2007). Despite being perceived as difficult, many EFL learners consider speaking to be the most important skill to develop (Kitano, 2001).

In order to help EFL learners develop greater speaking proficiency, various versions of communicative language teaching (CLT) have been implemented in English language classrooms in EFL contexts (Nunan, 2003; Savignon, 2005; Takanashi, 2004), as it offers learners the opportunity to develop speaking proficiency through authentic communicative tasks (Larsen-Freeman, 1986). Within the CLT framework, task-based language teaching (TBLT) has flourished in the past several decades and is being increasingly applied in EFL classrooms (Ellis, 2003, 2009b, 2018; Long, 2015; Nassaji & Fotos, 2011). Task-based language teaching gained wide recognition from second language acquisition (SLA) researchers in the 1980s because it offers learners the opportunity to develop speaking proficiency by engaging in activities that resemble real-

life situations. However, research has shown that although many learners become more fluent from completing communicative tasks, they continue to make numerous grammatical errors; thus, a major criticism of earlier task-based language teaching approaches (e.g., Long, 1991) is that they overemphasized the importance of meaning-focused instruction and lacked sufficient attention to form. Long's earlier definition of *focus on form* indicated that it is meaning-centered and incidental. Task-based language teaching has since progressed from its original emphasis on incidental and meaning centered learning to recognizing the importance of form-focused instruction (FFI) by integrating form into meaningful tasks (Ellis, 2009; Long, 2015). Form-focused instruction (FFI) refers to both incidental and planned instruction that directs learner attention to target language linguistic forms. The definition of form-focused instruction varies, as it can be both implicit and explicit, and can include pedagogical activities such as planning and repetition, teacher modeling, teacher explanations of target forms, and teacher feedback (Norris & Ortega, 2000).

However, the field of TBLT still lacks sufficient evidence for the positive effects of form-focused instruction. Up to now, a large number of researchers have examined the effects of form-focused instruction on speaking development using task repetition (e.g., Bygate, 1996, 1999, 2001; Bygate & Samuda, 2005; Kanda, 2015; Lambert, Kormos, & Minn, 2017) and pre-task planning (e.g., Ellis & Yuan, 2004; Foster & Skehan, 1996; Ortega, 1999; Yuan & Ellis, 2003) as pedagogical interventions. However, most previous studies concerning the effects of form-focused instruction have produced significant fluency gains (e.g., Ahmadian 2011; Ahmadian & Tavakoli, 2011; Bygate, 2001; Ellis,

2005a), some significant gains in syntactic complexity (e.g., Ahmadian, 2011; Bygate, 2001; Gass et al., 1999; Skehan, 2009a), but almost no study has produced significant gains in syntactic accuracy (see Gass, Mackey, Alvarez-Torres, & Fernández-García, 1999 for an exception). While previous researchers have acknowledged the positive effects of form-focused instruction on speaking development, more research on its influence on the development of syntactic accuracy is needed.

## Statement of the Problem

The first problem addressed in this study concerns the lack of investigations into the effects of form-focused instruction on speaking development. While TBLT researchers (e.g., Ellis, 2009a; Mochizuki & Ortega, 2008) have acknowledged the importance of integrating form-focused instruction and meaning-focused instruction for speaking proficiency development, further research is needed. To date, few studies concerning the effects of form-focused instruction on speaking development have produced syntactic accuracy gains (e.g., Fukuta, 2016; Kanda, 2015; Mochizuki & Ortega, 2008; Yuan & Ellis, 2003). A large number of form-focused instruction researchers have used repetition and pre-task planning as interventions and found positive results in the development of oral fluency (e.g., Ahmadian 2011; Ahmadian & Tavakoli, 2011; Bygate, 2001; Ellis, 2005a) and complexity (e.g., Ahmadian, 2011; Bygate, 2001; Gass et al., 1999; Skehan, 2009a) but very rarely in syntactic accuracy (see Gass et al., 1999, and Mochizuki & Ortega, 2008 for exceptions).

The second problem is the lack of longitudinal research in speaking proficiency development (Vercellotti, 2017). Few longitudinal studies of speaking development have been conducted, particularly with large sample sizes, because of the difficulty involved with the collection, transcription, and analysis of spoken data (Hughes, 2011). Due to the lack of longitudinal studies, it is unknown how learners develop their speaking proficiency over time when exposed to form-focused interventions.

The third problem addressed concerns the lack of understanding of the effects of form-focused instruction on learner affect. Affective factors are the least understood constructs in second language learning due to their abstract nature (Costa, 2015; Swain, 2013). The lack of understanding of learners' affective states toward speaking English is problematic because these states influence learners' cognition, behaviors, and learning experiences. While many studies have been conducted on learner affect and foreign language learning (Aida, 1994; Horwitz, Horwitz, & Cope, 1986; MacIntyre & Gardner, 1991; Mak, 2011), the research mainly involved correlational analyses of an affective variable and other variables (e.g., correlation between speaking anxiety and self-rated proficiency) or studies focused on the effects of affective variables and language achievement (Horwitz, 2010). Few studies concerning the effects of form-focused instruction on affective factors such as speaking anxiety, speaking self-efficacy, and desire to speak English, have been conducted (see Van Batenburg, Oostdam, Van Gelderen, Fukkink, & de Jong, 2019, and Graus & Coppen, 2016 for exceptions), in part because previous studies of form-focused instruction were mostly conducted from cognitive perspectives (Costa, 2015; Van Batenburg et al., 2019).

The fourth problem is that previous researchers have not investigated the effects of form-focused instruction on speaking proficiency from learners' perspectives. While SLA researchers have called for more research into the effects of form-focused instruction on task-based language learning, the results of previous studies have been based on objective, quantitative measurements. No researchers to my knowledge have investigated the effects of form-focused instruction on speaking development based on learner self-assessment; therefore, how learners perceive their speaking performance to have changed as a result of engaging in form-focused instruction remains unknown. This gap needs to be addressed due to the learner-centered nature of task-based language teaching; learners need to both collaborate and share responsibility for their own learning with their teachers (Nunan, 1988). In a learner-centered curriculum, learners would need to be included in the process of evaluating their own learning achievement (Little, 2005).

## Purposes and Significance of the Study

The first purpose of this study was to examine the effects of form-focused instruction on speaking development by looking at changes in syntactic complexity, syntactic accuracy, and oral fluency. In the CLT context, many teachers fail to distinguish between teaching specific linguistic features of the target language and teaching the target language by providing opportunities for learners to speak it. Communicative tasks alone appear to be insufficient to develop speaking proficiency; thus, the integration of explicit form-focused instruction into tasks needs to be investigated further. Form-focused instruction reduces the ambiguity of language learning by directing students' attention to

specific linguistic objectives in a structured environment through demonstrations, explanations, examples, and practice of specific linguistic forms. The importance of integrating form-focused instruction and task-based language teaching has been strongly supported by researchers such as Ellis (2006, 2009b, 2013, 2018), Long (2015), and Norris and Ortega (2000). Therefore, the first aim of this study is to contribute to the growing body of TBLT research. In particular, I aimed to examine changes in syntactic accuracy in speaking from the effect of form-focused instruction in terms of both global syntactic accuracy by using percentage of error-free T-units, as well as three past tense verb forms: simple past, past continuous, and past perfect. By focusing on these specific grammatical forms, it is possible to shed light on how the use of past tense in speaking develops through form-focused instruction over time. As such, it adds to our understanding of the development of syntactic accuracy in the above two senses.

The second purpose of this study was to examine the development of oral proficiency longitudinally with a considerably larger sample size than that used in past studies. Due to the lack of longitudinal speaking studies, the aim was to investigate the trajectory of changes in syntactic complexity, syntactic accuracy, and oral fluency over one Japanese academic semester. In this study, I investigated the speaking development of Japanese university students using a seven-week intervention from both descriptive and experimental perspectives in order to describe the dynamic nature of their oral language development.

The third purpose was to examine the effects of form-focused instruction on learner affect. Understanding learner affect is important because of the influential role it

plays in determining learner behaviors (Schunk, 1991). Research has shown that how learners feel about their language learning partly determine their learning experiences and outcomes. Sparks and Ganschow (1995) further argued that learner beliefs can be used to make a realistic assessment of language skills; therefore, this study offers another perspective for understanding speaking proficiency development by examining the effects of form-focused instruction on affective variables in terms of classroom English speaking anxiety, speaking self-efficacy, and desire to speak English.

The fourth purpose was to examine the effects of form-focused instruction on speaking proficiency development using learner self-assessment. In this study, the participants assessed their own speaking task performances in terms of syntactic complexity, syntactic accuracy, and oral fluency after receiving form-focused instruction. It is important for researchers and teachers to know how learners assess themselves because learners can make a meaningful contribution to their own evaluation. The growing interest in learner-centered teaching has placed greater focus on the use of learner self-assessment (Peirce, Swain, & Hart, 1993). Although past studies have produced controversial results and test administrators are often skeptical about using learner self-assessment to measure second language ability, past research is inconclusive because the importance of learner self-assessment cannot be measured by how strongly it correlates with other measurements but how it helps researchers and teachers to better understand learners' perceptions of their own language performance. Understanding learner self-assessment is important because of the learner-centered nature of task-based language teaching and the need to involve learners in the evaluation process (Little,

2005); therefore, in this study, I offer an alternative perspective for understanding the complex nature of speaking proficiency development.

## The Audience for the Study

The first audience for this study is SLA researchers, especially those interested in speaking and TBLT research. A large number of studies have been conducted in the field of speaking proficiency development, but more studies are needed to explore the dynamic nature of oral production. Speaking proficiency studies are comparatively underdeveloped compared to studies in listening, reading, and writing due to the processes involved with audio recording, transcribing, and data pruning. Therefore, the results of this study can add to the growing body of literature on speaking proficiency development. In addition, task-based language teaching has become a well-acknowledged approach and many researchers in the field of task-based language teaching have further acknowledged the importance of integrating form-focused instruction into communicative tasks. However, most studies up to date have only found significant gains in syntactic complexity and oral fluency, but not accuracy.

The second audience is classroom English teachers, especially those working in EFL contexts. Many EFL teachers struggle when teaching speaking courses (Hughes, 2011; Kramsch, 2014); therefore, teachers wanting to learn about effective approaches for developing their students' speaking proficiency can benefit from this study. This research provides data about pedagogical tasks that teachers can use when developing classroom tasks, lesson plans, and course syllabi. In addition, it provides insights into

learner affect—speaking anxiety, self-perceived efficacy, and desire to speak English. Teachers who want to understand how learners feel about foreign language learning as it relates to learners' learning experiences and outcome will be interested in the results of this study. Finally, teachers interested in learner self-assessment will benefit, as learner assessment often differs from teacher assessment. Understanding learner self-assessment is important for teachers in learner-centered classrooms given that they need to understand how learners can share responsibility and collaborate with teachers.

The third audience is policy makers, curriculum developers, and school administrators. The study provides valuable information regarding effective pedagogy for developing communicative speaking skills. This study provides new insights about English education by questioning the effectiveness of forms of communicative language teaching that eschew form-focused instruction. The Japanese Ministry of Education, Culture, Sports, Science, and Technology (MEXT) has endeavored to improve the speaking proficiency of Japanese students by promoting the implementation of a communicative language teaching curriculum for almost 25 years. However, the implementation of this policy has met with little success. This result suggests that there is a lack of understanding of communicative language teaching in Japan. In addition, communicative language teaching alone is not sufficient for EFL learners, as most learners appear unable to acquire much morpho-syntax from mere exposure to and communicative use of the target language. EFL learners need to be taught particular linguistic forms explicitly and given opportunities to use them communicatively. Despite criticisms of form-focused instruction as a traditional teacher-centered approach, it is

arguably an effective aspect of a bask-based pedagogy for EFL speaking development. The results of this study provide information concerning the effectiveness of combining form-focused and task-based instruction for future policy making, curriculum development, and school administrations.

## Delimitations

This study has five main delimitations. The first delimitation is that the participants are EFL learners who have few opportunities to use English communicatively outside of the classroom. Thus, the results should be applied with caution to ESL (English as a Second Language) learners, as they receive much more exposure to the target language than most EFL learners.

The second delimitation is the participants' first language, Japanese. Learners' first language greatly influences the development of second language competence. Japanese and English are two languages that have very different syntactic structures; therefore, the results of this study cannot be applied directly to learners with a different L1, particularly to learners whose L1 is genetically related to English.

Third, the participants have particular characteristics in terms of academic ability, attitude toward learning, memory capacity, and second language proficiency. The present study was conducted at a prestigious national university in Japan where the students had passed highly competitive entrance examinations. Therefore, they are considered more academically capable, more motivated to learn, and to have higher English language proficiency than typical Japanese university students. Therefore, the same research might

generate different results if it involves students with different academic aptitude, proficiency levels, and background.

Fourth, the form-focused intervention in this study was limited to three past tense forms—simple past, past continuous, and past perfect—which are considered salient and relatively simple. Therefore, it is unclear if the results of this study would apply to more complex grammatical forms.

Fifth, the primary spoken data collected in this study were derived from monologues; therefore, the results cannot be applied directly to dialogic interactions because monologue development differs from dialogue development and because of the lack of interaction in monologues.

## Organization of the Study

In Chapter 2, Review of the Literature, I review past studies related to speaking development in the following sections: conceptualizing speaking, conceptualizing second language speaking proficiency, measurement of second language speaking proficiency, communicative language teaching, task-based language teaching, and learner affect toward English speaking. At the end of Chapter 2, I present the gaps in the literature, the purposes of the study, and the research hypotheses that guide the present study. In Chapter 3, Methods, I describe the research setting, participants, selection of target forms, intervention groups and the comparison group, instrumentation, procedures, data deletion, data analyzing procedures, and analysis of the research hypotheses. In Chapter 4, Preliminary Analyses, I report the initial findings from the English Speaking Learner

Affect Questionnaire and Self-Assessment of Speaking Task Questionnaire using Rasch model analysis, initial findings from the analytical rating using multifaceted Rasch analysis, and the frequency of three types of part tense forms observed. In Chapter 5, Results, I present results for the five research hypotheses. In Chapter 6, Discussion, I discuss the results and present the pedagogical implications. In Chapter 7, Conclusion, I summarize the findings, present the limitations of the study, suggestions for future research, and conclude with final remarks.

# CHAPTER 2

# REVIEW OF THE LITERATURE

This chapter has four parts. First, I conceptualize speaking and second language speaking proficiency by examining three constructs: syntactic complexity, syntactic accuracy, and oral fluency. Second, I review the literature related to the measurement of second language speaking proficiency by looking at CAF measurements, analytical ratings by human raters, and learner self-assessment. Third, I review the development of speaking proficiency in using task-based language teaching before reviewing the effects of form-focused instruction and task complexity. Fourth, I review three affective variables related to English speaking—classroom English speaking anxiety, speaking self-efficacy, and desire to speak English. After reviewing the previous literature, I describe the gaps in the literature, the purposes of the study, and the research hypotheses that guide the present study.

## Conceptualizing Speech Production

Speaking is a unique form of communication only possible to human beings. A voice provides speakers with a way to immediately project their self to the world, which makes speaking more communicative yet less tangible than reading, writing, and listening (Hughes, 2011). Speaking is the most primary form of communication, as children learn to speak before they learn to write or read. In the first language (L1) context, children acquire speaking ability by interacting with adults and other children in

the surrounding environment. Adult speakers also create opportunities for children to use language to express themselves and to interpret utterances by others, and they assist children's speaking development through modeling, reformulation, and clarification (Goh & Burns, 2012). Speaking in the first language is not effortless, as it involves the interaction of complex cognitive processes that demand large amounts of linguistic knowledge and working-memory processing capacity (Goh & Burns, 2012).

First, language speech production consists of four components: conceptualization, formulation, speech articulation, and self-monitoring (Bygate, 2009; Kormos, 2005; Levelt, 1989). Conceptualization deals with planning what to say, formulation deals with the grammatical, lexical, and phonological encoding of message, articulation deals with the production of actual oral sounds, and self-monitoring deals with checking the accurateness of the produced oral outputs (Kormos, 2005). There are two L1 speech production models: spreading activation (e.g., Dell, 1986) and modular (e.g., Levelt, 1989). While both models are similar in that they assume that information is transmitted in the speech-processing model through activation, the spreading activation model assumes that activation can go both forward and backward, while the modular model assumes that activation only moves forward (Kormos, 2005). According to the spreading activation model, a warning signal is produced when an error occurs at any particular level so that activation flows up to the next level and processing starts again at the next level. The modular model is similar to the spreading activation model but the error does not get noticed so warming signal is be produced at its current level. As the result, errors can only be noticed when they are being phonologically encoded or when the utterance is

articulated. Therefore, according to the modular model of production, processing must start from the conceptualization stage, as this is the first place where erroneous information can originate. Another difference between the two models is that the spreading activation model has the assumption that speakers construct frames for sentences and phonetic representations first, and then select the appropriate pieces of information to fit in the constructed frames, whereas the modular model assumes that speakers construct words that activate syntactic building, which then activates phonological encoding (Kormos, 2005).

Levelt's modular speech production model indicates that speech production happens through processing in the conceptualization, formulation, and articulation components (Kormos, 2005). A component only starts processing after it receives input from the previous component. However, parallel processing can take place, as the conceptualizer can start working on the next piece of information while the previous piece of information is still being processed. Therefore, the articulation of a sentence can begin before the speaker has completed the planning of the entire message because of the automatous and parallel nature of processing. According to Figure 1, speech production begins from conceptual preparation where the message is planned. There are two types of planning—macroplanning and microplanning—occurring at the conceptual preparation stage. Macroplanning involves the planning of communicative intention such as informing, directing, requesting, and apologizing, whereas microplanning involves deciding on the perspective the speaker takes to convey the message. Therefore, the microplanning stage also determines whether a noun or a pronoun is used as well as

giving propositional content to the message, such as specifying the argument structure of the message, specifying the referent, and determining the mood of the message. The microplanning stage also involves the selection of language-specific information such as the selection of verb tense(s). Conceptual preparation leads to preverbal messages, which contain all the necessary information to convert meaning into language. Preverbal messages are the input for grammatical encoding, where the selection of lexical units and syntactic encoding takes place by retrieving information from the speaker's mental lexicon storage. Lemma activation occurs in the grammatical encoding phase where the selection of a lemma activates the syntactic structure building procedures. The output of grammatical encoding becomes processed into morpho-phonological encoding where the morphemes constituting the word are accessed and features such as stress and pitch are set. This process leads to the phonological score, which provides the input for phonetic encoding, and the phonetic encoding generates the articulatory score when it is converted into overt speech through articulation. In this production model, monitoring happens throughout all of the stages, but it can be categorized into three loops. The first loop monitors speech production by comparing the preverbal plan with the original intention, the second loop monitors internal speech before articulation, and the third loop happens when the utterance is checked for accurateness after articulation. Upon perceiving an error in the output in any of these three loops, the monitor issues an alarm signal, which triggers the production mechanism for a second time (Kormos, 2005).

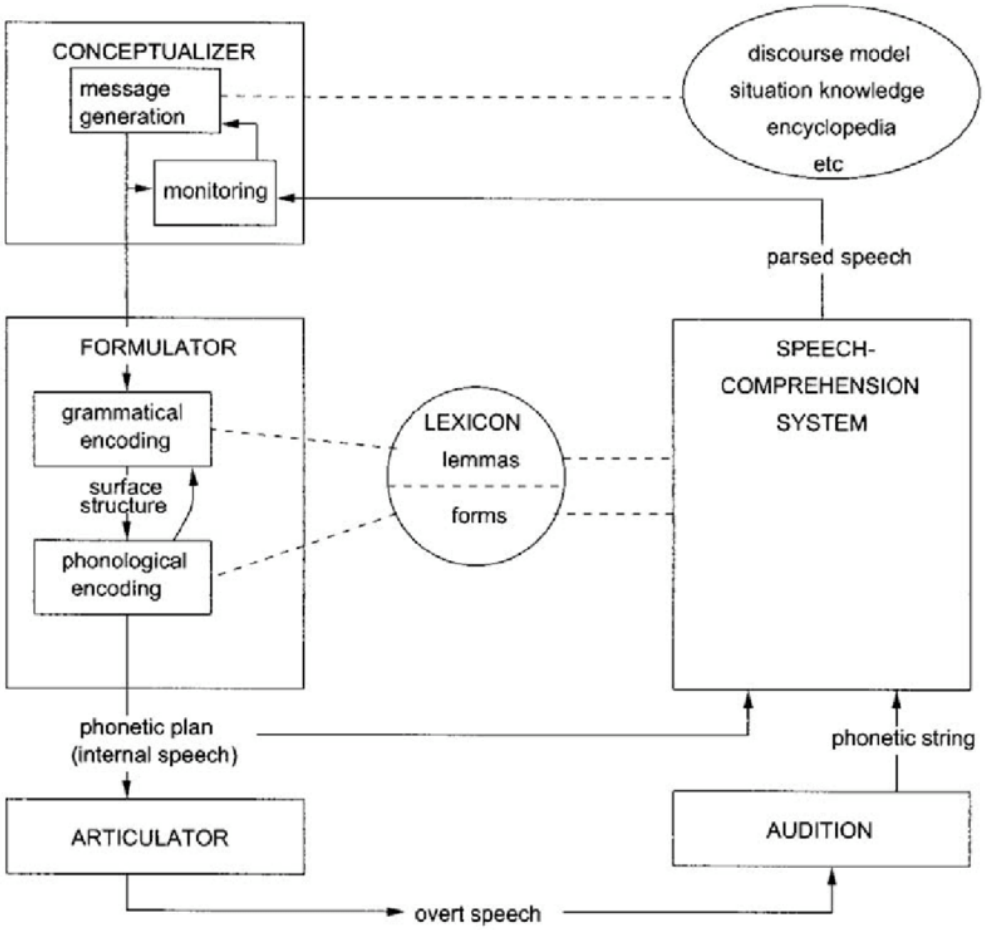



*Figure 1*. Blueprint of speech production (Levelt, 1989).

In second language speech production, Levelt's (1989) modular model has been most widely referenced (Bygate, 1998, 2001; de Bot 1992; Ellis 2009a; Kormos, 2005). However, it has been argued that while Levelt's model has been widely referenced, it cannot be used directly with L2 speech production because of the fundamental differences in the size, elaborateness, and organization of L1 and L2 lexicons (Skehan, 2009b). Speaking in a second language is understandably more difficult than speaking in a first language because L2 linguistic knowledge is often missing, partial, or inaccurate. In addition, speaking in an L2 places more demands on working memory because

accessing, proceeding and interacting multiple cognitive processes is generally less automatic compared to speaking an L1. As a result, L2 speech production is often less fluent, accurate, and complex than first language speech production.

Similar to Levelt's (1989) first language model, the bilingual speech production model also involves the four processes of conceptualization, utterance formulation, speech articulation, and self-monitoring (de Bot, 1992; Kormos, 2005). Kormos' (2005) model of bilingual speech production builds on Levelt's first language model (see Figure 2). The bilingual speech production model also assumes that speech production occurs in a modular manner where the present component completes and activates the next component; however, the bilingual speech model is slightly different where it is not completely serial because cascading activation happens between the lexical and phonological levels of encoding. Most stored knowledge is shared between L1 and L2 items, but second language production also involves the use of stored declarative knowledge of syntactic and phonological rules (Kormos, 2005). Another major difference between the two models is that the bilingual speech production model assumes that there is competition between L1 and L2 items and encoding procedures and the need for a compensatory mechanism in order to make up for missing knowledge in L2 production. For L2 speakers, formulation can be the most difficult stage because it involves making lexico-grammatical choices and deciding which words and morpho-syntactic form combinations are most appropriate for conveying the intended messages. Speakers accomplish this task by relying on their knowledge of the target language lexico-grammatical system (Goh & Burns, 2012). In addition to lexical and morphosyntactic

issues, L2 speakers must also attend to coherence and discourse organization, activate phonological forms, retrieve articulatory gestures, and monitor themselves throughout the speech production process (Revesz, Sachs, & Hama, 2014).

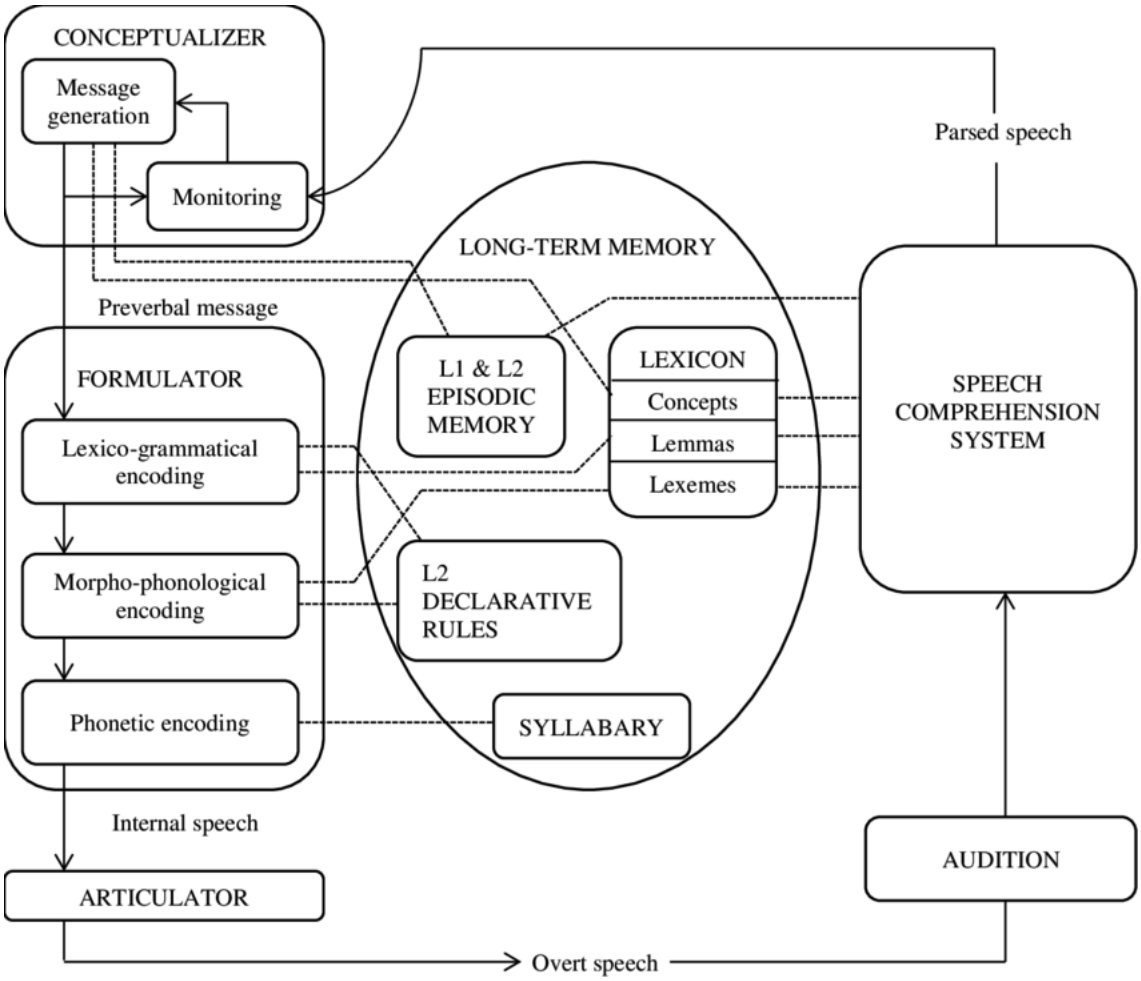



*Figure 2*. The model of bilingual speech production (Kormos, 2005).

Monitoring is considered a major L2 speech production process (Kormos, 2005) because it allows learners to notice deficiencies in their knowledge of the target language (Izumi, 2003). While different models have been used to explain the different roles of

monitoring in speech production, overall, monitoring in L2 speech production is similar to L1 speech production where it happens in three major phases: when learners monitor speech production by comparing the preverbal plan with the original intention, monitor the internal speech before articulation, and monitor after articulation when the produced oral utterance is checked for accurateness (Kormos, 2005). Second language monitoring is closely related with self-repairs (Kormos, 2005). However, self-initiated self-repairs are only an overt production of the monitoring processes when learners identify an output that is inappropriate or erroneous. However, monitoring happens before oral articulation and speakers either repair the parts before an utterance or decide not to correct the mistakes in the utterance. The speed of self-repair varies but generally, it takes less time for learners to repair on small phonological issues than to repair on large-scale changes.

Simply speaking the target language without consciousness is inadequate for speaking development to occur (Goh & Burns, 2012; Skehan, Bui, & Wang, 2012). Second language learners need comprehensible output where they develop speaking by being pushed to produce precise and appropriate output (Swain, 1985). According to the Comprehensible Output Hypothesis, comprehensible output refers to improved output in terms of improved content or other grammatical, sociolinguistic, or discourse features (Swain, 1995). Therefore, producing the target language appropriately and precisely triggers learners to pay attention to the language needed to successfully convey their intended meaning (Swain, 1995). It is also through comprehensible output that learners are pushed to move from semantic to syntactic processing as they are able to further analyze the grammar in the target language. Swain (1995, 2005) further proposed that

comprehensible output has a noticing/trigger function. According to the noticing/trigger function, learners can notice what they are lacking when attempting to produce output precisely and appropriately. Therefore, the act of producing the L2 can prompt second language learners to recognize their own linguistic gaps. This awareness is the mechanism that triggers the cognitive processes needed for second language learning.

According to the Noticing Hypothesis, input does not turn into intake for language learning unless it is consciously noticed and learners pay attention to it (Schmidt, 1990, 2001). Learners learn about things they pay attention to and learn little or nothing about the things they do not notice. Even forms that occur frequently in the input, are not acquired until they are consciously noticed. For this reason, foreign language learners can have limited grammatical development while successfully developing L2 communicative competence. Adult second language learners need to pay conscious attention to linguistic form and reflect on linguistic structure. In addition, learners' grammatical errors can be frequently corrected, but the corrections have little effect if learners are unaware that they are being corrected (Schmidt & Frota, 1986). The Noticing Hypothesis (Schmidt, 1990, 2001) is related to Noticing the Gap Hypothesis (Schmidt & Frota, 1986) where learners only overcome errors if they make a conscious comparison between their output and the target language input. By noticing inappropriate language forms in their own speech and modifying their spoken language to achieve more target-like performances, learners develop grammatical accuracy.

## Conceptualizing Second Language Speaking Proficiency

Because second language speaking proficiency is a multi-dimensional concept, past researchers have proposed varying operational definitions (Ellis, 2009a; Hughes, 2011). It is also conceptualized differently depending on the discourse that it is being analyzed such as pragmatics, linguistic, functional, interactional, conversational, and sociocultural discourses (Ellis & Barkhuizen, 2005). The conceptualization of L2 speaking proficiency is controversial because it is not always clear what speaking proficiency entails, and this lack of clarity makes the measurement of speaking proficiency difficult and also complicates comparisons of multiple studies (Ellis, 2009a; Housen & Kuiken, 2009; Housen, Kuiken, & Vedder, 2012; Iwashita, 2010). From a linguistic perspective, researchers in the past three decades have often conceptualized second language speaking proficiency in terms of three constructs: complexity, accuracy, and fluency (Ellis & Barkhuizen, 2005; Norris & Ortega, 2009). This CAF triad was proposed in the 1990's after Skehan (1989) added the complexity construct to the previously conceptualized constructs of fluency and accuracy Brumfit (1984).

Second language speaking proficiency has been measured using various methods ranging from qualitative, subjective measurements (e.g., Adams, 1980) to quantitative, objective measurements. (e.g., Ahmadian, 2011; de Jong & Vercellotti, 2016; Mochizuki & Ortega, 2008). Different researchers have used different measures to determine the degree to which oral production is complex, accurate, and fluent, so unsurprisingly, there is limited agreement as to which measurements are most suitable for assessing speaking proficiency because they are defined differently depending on the learning purposes,

learners and contexts (Housen & Kuiken, 2009; Norris & Ortega, 2009). However, different measurement and analytical methods can produce different research results (Nitta & Nakatsuhara, 2014; Norris & Ortega, 2003; Ortega, 2003). While measurements of speaking proficiency vary, most researchers (e.g., Norris & Ortega, 2009) have agreed that speaking performances need to be measured using multiple measures in order to gain more precise and objective accounts of oral proficiency. Researchers who have used single dimension measurements (e.g., only fluency measurements) have reported more positive results from the effects of intervention than those who have used multi-dimensional measurements of complexity, accuracy, and fluency measurements together (Bygate, 1996). For example, researchers who measured speaking proficiency development using fluency measures have generally produced positive results through their interventions (e.g., Ahmadian 2011; Ahmadian & Tavakoli, 2011; Bygate, 2001; Ellis, 2005a), whereas those who used syntactic complexity measures have reported both positive and negative results from the effects of interventions (e.g., Larsen-Freeman, 2006; Skehan, 2009a). Finally, researchers who measured the development of speaking proficiency using syntactic accuracy measures have very rarely produced positive results (see Gass, Mackey, Alvarez-Torres, & Fernández-García, 1999 for an exception).

Foster, Tonkyn, and Wigglesworth (2000) addressed the issue of speaking proficiency measurement by examining the measurements used in studies in four leading SLA journals: *Applied Linguistics*, *Language Learning*, *Studies in Second Language Acquisition*, and *TESOL Quarterly*. They found that the biggest problem surrounding the measurement of speaking proficiency is the lack of agreed upon definitions and clear

explanations of measurement units. Foster et al. (2000) further reported that of 87 studies they examined, 44 included no description of the measurement units used. The other 43 studies included definitions of the measurement units but different definitions were used for the same measurement units.

### Syntactic Complexity

Complexity, which is the extent to which target language production is elaborated and varied (Ellis, 2003), can be divided into cognitive complexity and linguistic complexity (Housen et al., 2012). Complexity is also considered the most controversial dimension of the CAF triad (Michel, 2017). Cognitive complexity is learner dependent and concerns features such as aptitude, memory span, motivation, and L1 background, whereas linguistic complexity, which defines learners' interlanguage system in terms of size, elaborateness, richness, and diversity of language production (DeKeyser, 1998), can be divided into four dimensions: syntactic (sentential, clausal, and phrasal levels), lexical (word and collocation levels), morphological (inflectional and derivational levels), and phonological (segmental and suprasegemental levels) complexity (Koizumi & In'nami, 2014). Unlike Koizumi and In'nami, who classified lexical complexity as a part of the overall complexity construct, some researchers have argued that lexical performance should be a fourth construct in addition to complexity, accuracy, and fluency in order to more completely conceptualize the nature of L2 speaking proficiency (e.g., Malvern & Richards, 2002; Skehan, 2009a). However, the addition of lexical performance as the fourth construct on top of the existing CAF triad is also controversial because some

studies (e.g., Skehan, 2009a) also argued that lexis does not need to be included as the fourth construct because the effects of lexical performance on L2 speaking proficiency are not as significant as they are on L1 speaking proficiency. Up to date, most task-based language learning researchers investigating the development of complexity have focused on its syntactic component (Michel, 2017; Norris & Ortega, 2009).

Syntactic complexity has been measured across a wide range of language-related fields using multiple measurements (Norris & Ortega, 2009; Yuan & Ellis, 2003). Most researchers measure syntactic complexity using multiple measurements because one component of complexity (e.g., subordination complexity) can stabilize while another component (e.g., global complexity) continues to develop (Scott, 1988). In addition, functional linguists have argued that learners develop from (a) the expression of ideas by means of coordinating or sequencing single words, sentences, and clauses to (b) an expansion by which subordination is added as a resource to expressing logical connections of ideas, and finally to (c) the emergence of and reliance on grammatical metaphor, which leads to advanced language with greater lexical density, which is a measure of the proportion of lexical items (i.e., the proportion of nouns, verbs, adjectives, adverbs used), and more complex phrases (Halliday & Matthiessen, 1999). Therefore, to capture the development of complexity across these processes, three sub-dimensions need to be measured: (a) global complexity, (b) complexity by subordination, and (c) complexity via sub-clausal or phrasal elaboration (Norris & Ortega, 2009).

Global complexity is often measured by mean length of T-unit (Norris & Ortega, 2009). It can be further measured using 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 (Mochizuki & Ortega, 2008). In narrative monologues, T-units are often used as a measurement unit (Yuan & Ellis, 2003). A T-unit is defined as a main clause with or without subordinate clauses (Mochizuki & Ortega, 2008). This definition is similar to Hunt's definition of a T-unit, which was defined as one main clause plus subordinate clauses attached or embedded in it (Hunt, 1970). In addition to the T-unit, the C-unit and AS-unit are also used for analyzing speaking data because they are considered appropriate for analyzing oral data containing ungrammatical segments (Foster et al., 2000; Norris & Ortega, 2009). Finally, C-units are considered to be suitable for analyzing lower-proficiency learners, whereas T-units are considered suitable for intermediate-level learners (Norris & Ortega, 2009). Complexity by subordination can be calculated by counting all clauses and dividing them over a given production unit (e.g., clauses per T-unit or clauses per C-unit or clauses per AS-unit) (Elder & Iwashita, 2005). Finally, complexity by sub-clausal or phrasal elaboration is measured using mean length of clause (Scott, 1988). However, while it is important to measure syntactic complexity multidimensionally, it has been argued that there are often redundancies in the syntactic complexity metrics, as some measurements might be measuring the same thing (Norris & Ortega, 2009).

**Syntactic Accuracy**

Accuracy is considered the most straightforward construct in the CAF triad, as it is a measure of the target-like and error-free use of language (Housen & Kuiken, 2009;

Michel, 2017; Mochizuki & Ortega, 2008). However, while it is simply a measure of deviations from native-speaker norms (Housen & Kuiken, 2009), there are controversies surrounding what criteria are used when deciding on the choice of norms and also how far away the deviations are from the chosen norms (Michel, 2017).

Syntactic accuracy can be measured in terms of global accuracy (Foster & Skehan, 1996; Iwashita et al., 2008; Skehan & Foster, 1999) and the accuracy of specific linguistic forms (Iwashita et al., 2008; Ortega, 1999; Wigglesworth, 1997). Global accuracy, which is viewed as the most comprehensive of all accuracy measures because all errors are considered, is frequently measured by the percentage of error-free T-units or percentage of error-free clauses (Michel, 2017). The percentage of error free T-units refers to T-units that are free from grammatical errors including both specific errors (e.g., past tense) and other common linguistic errors such as word order errors, morphological errors, and the omission of obligatory forms such as pronouns (Iwashita et al., 2008). However, while global accuracy can be a holistic measure that includes all deviations from the target language norms, it is not sensitive enough to capture small differences in learners with higher levels of proficiency or changes from short-term interventions (Lambert & Kormos, 2014). On the other hand, analyses of specific types of errors have been conducted on features such as verb tense, third person singular, plural markers, prepositions, and articles (Wigglesworth, 1997). The choice for a specific accuracy measure is based on the language that is expected. For example, when investigating the effects of teaching past tense, target-like-use of past tense *-ed* would be chosen as the target specific measure (Michel, 2017). However, while specific error coding can offer a

detailed description of erroneous target forms, they cannot represent holistic accuracy performance in learners (Iwashita et al., 2008). In addition, it is difficult to generalize the findings of specific measures to other contexts (Michel, 2017).

## Oral Fluency

Oral fluency refers to the smoothness, rapidness, effortless use of language. Such language is characterized by a limited number of pauses, hesitations, and reformulations (Hilton, 2008; Koponen & Riggenbach, 2000; Lennon, 1990, 2000; Michel, 2017). Automaticity is a key component of oral fluency; fluent speakers can more speedily retrieve lexical items from memory, encode grammatical forms, and correct their own erroneous output than less fluent speakers (Segalowitz, 2003, 2010). While fluency is a multi-dimensional construct (Lennon, 2000), it is considered to be relatively uncontroversial compared to complexity and accuracy (Michel, 2017). Fluency has received a large amount of research attention in past studies of oral proficiency development because it is the most salient construct and fluent speakers are often considered to be successful language learners (e.g., Iwashita, Brown, McNamara, & O'Hagan, 2008). However, some aspects of oral fluency have been found to be more closely related to personal traits than indicators of language competence (de Jong, Groenhout, Schoonen, & Hulstijn, 2015).

There are many different methods to measure oral fluency, but it is commonly measured by speed fluency, repair fluency, and breakdown fluency (Freed, 2000; Mora, 2006; Skehan, 2003; Towell, 2002). Speed fluency is associated with control and access

to proceduralized knowledge (Michel, 2017) and is usually measured by time-related measures such as speech rate and mean length of runs (Kormos & Dénes, 2004). Speech rate refers to the number of syllables produced per minute while articulation rate refers to the number of syllables produced per minute after excluding pause time (Tavakoli & Skehan, 2005). Speech rate is considered a valid quantitative measure because it takes varying word lengths into account (Kormos & Dénes, 2004; Mora, 2006). Mean length of runs refers to the average number of syllables between pauses (de Jong, 2016). Despite debates in the literature regarding what the cut-off point of pause length should be, it is commonly accepted that mean length of run refers to the number of syllables produced in utterances between pauses of 0.25 seconds and above (Kormos & Dénes, 2004; Lennon, 1990; Towell, Hawkins, & Bazergui, 1996). Repair fluency, which is seen as an indicator of monitoring processes (Levelt, 1989, Kormos, 2005), is usually measured by dysfluency markers such as reformulations, repetitions, false starts, and replacements (Skehan, 2003). Breakdown fluency is usually measured by silence-related metrics such as pausing, which can be divided into unfilled and filled pauses such as *mmm*, *eeeh*, and *aaah*. Second language speakers pause more often within clauses than L1 speakers (Skehan & Foster, 2007). Pausing fluency is difficult to examine because it is considered to be more related with L1 proficiency than L2 proficiency (de Jong et al., 2015).

Mean length of pauses is calculated by dividing the total length of pauses above 0.2 seconds by the total number of pauses above 0.2 seconds (Towell et al., 1996). The measurement of unfilled pauses is complicated by the fact that different researchers have defined an unfilled pause as ranging from 0.28 to 3.0 seconds (Riggenbach, 1991;

Towell, 2002). The role of filled and unfilled pauses in measuring the fluency construct is controversial as some previous studies with smaller number participants have indicated that measuring the amount of filled and unfilled pauses can help distinguish fluent speakers from non-fluent speakers (Freed, 2000; Lennon, 1990; Riggenbach, 1991), whereas other studies with larger number of participants found that filled and unfilled pauses do not correlate with overall ratings for oral fluency (Kormos & Dénes, 2004; van Gelderen, 1994). Furthermore, disfluencies tend to occur in clusters in the speech of non-fluent L2 speakers (e.g. Freed, 2000; Riggenbach 1991), while fluent speakers tend to pause at grammatical junctures (Lennon, 1990; Towell et al., 1996).

## The Dynamic Relationship of Complexity, Accuracy, and Fluency

While it has been widely acknowledged that proficient L2 speakers have the ability to use complex structures, speak accurately, and fluently (Ellis, 2009a), there is no consensus on the effects of the interaction among these three constructs on speaking proficiency (Housen & Kuiken, 2009; Skehan, 2009b) nor on the weighting of these constructs when determining how speaking proficiency is conceptualized (Larsen-Freeman, 2009). Learners can produce fluent but grammatically inaccurate speech, or speak fluently but lack a varied range of sentence structures or vocabulary, or speak accurately but not fluently (Lennon, 1990). The interaction between the CAF triad becomes more controversial due to the complex subsystems within each construct (Michel, 2017). It is considered impossible to measure one isolated single construct as all three constructs are interrelated; for example, grammatical knowledge correlates strongly

with fluency measures (Hilton, 2008; Sato, 2014). Accuracy and fluency are positively related because fluent speech entails the application of accurate processing mechanisms (Kormos & Dénes, 2004).

In addition to the complex interrelations between complexity, accuracy, and fluency, researchers have argued that trade-off effects between the three constructs could occur due to limited working memory capacity (Bygate, 1998; Ellis, 2005a, 2009a; Foster & Skehan, 1996) as well as the increased cognitive demands that comes with improved accuracy and complexity (Robinson, 2001). To date, most studies (e.g., Ahmadian, 2011; Ahmadian & Tavakoli, 2011; Yuan & Ellis, 2003) have pointed to trade-off effects between the three constructs. According to the Trade-off Hypothesis (Skehan, 1998), attentional capacity and working memory are limited; therefore, spending more attentional resources on one aspect negatively impacts one or both of the others. For example, speakers who prioritize accuracy tend to use the forms they have internalized and therefore become resistant to using more complex and less familiar forms (Skehan, 2009a). Complexity and accuracy almost never improve together as accuracy competes with complexity (Housen & Kuiken, 2009). Speakers can compensate for inadequacy in one area by drawing on the two other areas (Ellis, 2009a). In contrast to the Trade-off Hypothesis, Robinson (2001) proposed the Cognition Hypothesis, which suggests that the three constructs could improve together without trading off with one another, as long there are sufficient opportunities for input, practice, and output. Robinson further argued that increased cognitive demands can increase learner noticing and allow for learners the opportunities to rehearse in their memory, which can lead to the incorporation of new

linguistic forms and modifications in output. Housen and Kuiken (2009) summarized this debate by suggesting that it is possible for the three constructs to display trade-off effects in some contexts but improve together in other contexts when an appropriate cognitive demand is imposed.

Due to the dynamic nature of L2 language development (Larsen-Freeman, 2006), the understanding of the development of multiple subsystems and their relationship with each other, is more difficult to measure than any static single construct. For example, syntactic complexity does not always develop linearly in the dynamic process of L2 development (Lambert & Kormos, 2014; Larsen-Freeman, 2009).

## Task Complexity

Cognitive demand refers to the amount of attention, memory, reasoning, and other information processing required during a task (Robinson, 2001). As such, task complexity influences language learning and performance outcomes because different degrees of complexity pose different cognitive demands that influence task performance (Sasayama, 2016). Robinson proposed that task performances can change because the cognitive demands of tasks change depending on factors such as the number of task elements, reasoning demands of the task, immediacy of information provided, planning time, and prior knowledge (Robinson, 2001, 2011). Robinson further argued that tasks with greater cognitive demands could facilitate complexity and accuracy but decrease oral fluency. On the other hand, the effects of task complexity on task performance have been viewed differently by Skehan (1998, 2014), who suggested that increased task

complexity requires the use of more attentional resources, therefore making it more difficult for learners to pay attention to multiple aspects of the language simultaneously. Skehan and his colleagues (e.g., Foster & Skehan, 1996; Skehan & Foster, 1997, 1999) have reported that an increase in task complexity leads to decreased syntactic accuracy and oral fluency, and a decrease in task complexity leads to increased fluency or accuracy, but not both (Skehan, 1998). Learners trade off accuracy and fluency with more complex tasks due to limited attentional resources (Skehan, 2009a). However, other researchers have argued that while task complexity influences oral performances, it does not affect monologues as much as dialogues (Michel, Kuiken, & Vedder, 2007). Skehan (1998) proposed three categories of factors that determine task complexity: code complexity (e.g., linguistic complexity/variety and vocabulary complexity/variety), cognitive complexity (e.g., familiarity of task topic, information, and organizational structure), and communicative stress (e.g., time pressure, the nature of the prompt, and the number of participants).

In storytelling tasks, task complexity can be determined by the storyline and indirect messages embedded in the story that differ from the storyline portrayed by the illustrations (Sasayama, 2016). Indirect messages in the storyline are messages not portrayed directly by the illustration; thus, learners must interpret the indirect messages using past experience or knowledge. The complexity of storytelling tasks also depends on the topic, structure, and picture quality (Elder, Iwashita, & McNamara, 2002). In a speaking narration study conducted by Sasayama comparing the complexity of four tasks, the participants indicated that the storyline of one particular task was relatively complex

because they found it difficult to understand a twist in the story and the message the author wished to convey; therefore, the participants had to pay attention to small details in the pictures to decipher the indirect message of the story. The results indicated that the participants found storylines in which the characters interacted to be more difficult than those in which the characters acted individually because in the former case they had to not only explain each character's actions, but also explain their relationship to each other. The difficulty in deciphering the details of the pictures raised the cognitive demands of the task. Additionally, language-related issues, especially not knowing vocabulary needed to describe the story, made the task more complex for some participants. Therefore, even with equivalent tasks (i.e., four-picture cartoons at a particular difficulty level), there are unanticipated dimensions of complexity embedded in each task (Elder et al., 2002). It is difficult to produce tasks at exactly the same level of complexity even when the tasks have the same task characteristics and conditions (Brindley & Slatyer, 2002). Therefore, tasks hypothesized to be equivalent, need to be piloted and calibrated for complexity levels to produce equivalent task demands (Brindley, 2001).

## Analytical Ratings

In addition to CAF measurements aimed at measuring performance outcomes in terms of complexity, accuracy, and fluency, oral proficiency is also often assessed by human raters using rubrics (Cai, 2015; Kuiken & Vedder, 2014; Sato, 2011). While the relationship between rater assessment and CAF measurements is still unclear (Brown, 2000; Nitta & Nakatsuhara, 2014); rater assessment is a crucial measurement component

because raters offer an impressionistic judgment of the overall quality of oral performance (Nitta & Nakatsuhara, 2014). Two types of ratings are holistic and analytical ratings. While holistic ratings can describe overall oral performance by generating one score, the interpretation of holistic rating can be ambiguous because the weighting of the subcomponents that exist within the holistic score is not specified. Holistic ratings are more controversial because different raters implicitly weight the subcomponents of speaking differently (Xi & Mollaun, 2006) given that different raters have different perceptions of how particular subcomponents impact overall speaking proficiency (Brown, 1995). For example, Sawaki (2007) found that grammatical accuracy was the largest determinant of overall speaking performance because the teachers in her study focused mostly on grammatical errors.

The first advantage of analytical ratings is that they offer rich information about specific aspects of speaking proficiency (Brown & Bailey, 1984; Sato, 2011). Analytical ratings also offer a more precise description of oral proficiency because one construct can have overriding effects on other constructs. For example, accuracy "overrides the effect of temporal factors on listeners" because learners who are grammatically accurate can be perceived as more fluent than learners who speak faster but with lower grammatical accuracy (Kormos & Dénes, 2004, p. 160). Second, analytical ratings provide diagnostic information for teachers so that they can better understand students' strengths and weaknesses (Brown & Bailey, 1984). In a classroom context, analytical ratings are a useful tool that allows teachers to provide consistent and precise feedback to students. It offers teachers the ability to precisely detect learners' progress over time in individual

constructs. Third, analytical rating schemes allow test designers the opportunity to decide what specific components of speaking to assess because it is virtually impossible to assess all the potential components in a single test (Wigglesworth, 2008). Using analytical ratings, raters can distinguish target linguistic components from overall oral communicative ability. In addition, analytical ratings increase rating accuracy by focusing rater attention on specific rating criteria, and it is well aligned with the multi-componential nature of language proficiency (Bachman, Lynch, & Mason, 1995). Fourth, analytical ratings are useful for assessing learners who have uneven profiles across different aspects of speaking. For example, some learners have excellent fluency in terms of speed, repairs and breakdown fluency, but have much lower grammatical accuracy; others have an excellent control of sentence structure, but do not know how to organize their speech logically. Finally, analytical scoring is suitable for inexperienced raters as they can evaluate specific criteria, and the multi-dimensional criteria also make it easy for training and standardization purposes (Weir, 2005).

However, analytical scoring also has problems in terms of implementation and implication. First, it imposes a higher cognitive load on raters as they strive to assess multiple aspects of speaking proficiency; this situation can lead to rating inconsistency and rater fatigue (Xi & Mollaun, 2006). Second, raters have been criticized for subjectivity because they often do not agree upon the precise definitions of the analytical subcomponents. The subcomponents from the analytical scores can still have unequal weightings among raters, as it is difficult for raters to achieve a consensus regarding how speaking proficiency is conceptualized and the weighting of subcomponents (Larsen-

Freeman, 2009; Sawaki, 2007). Therefore, it is important to rate multiple aspects of oral proficiency. Third, the outcome of analytical ratings is influenced by factors such as task type, task complexity, task content (Sato, 2011), and rater characteristics (e.g., training, background, experience, and fatigue) (Davis, 2016; Ling, Mollaun, & Xi, 2014; Xi & Mollaun, 2006). For example, some raters might rate fluency more accurately than accuracy and complexity because of past rating experiences. Fourth, the results of analytical ratings can be influenced by the content of the speaking task even when content is not included as a subcomponent for analytical rating (Sato, 2011). Analytical ratings are usually based on assessing the linguistic components of the oral performance, which implicitly implies that oral performances are a representative of linguistic ability.

## Learner Self-Assessment

Learner self-assessment is a type of alternative assessment that differs from traditional teacher-oriented measurements (Alderson & Banerjee, 2001; Brown & Hudson, 1998). While many past researchers have found that learner self-assessment does not predict actual oral proficiency, it promotes reflection and increases learners' awareness, motivation, investment, and agency (Ma & Winke, 2019). Despite its reputation as an informal, subjective form of assessment (Peirce et al., 1993), learner self-assessment has gained research attention due to a growing recognition of its pedagogical and practical value (Alderson, 2005; Trofimovich, Isaacs, Kennedy, Saito, & Crowther, 2016). First, self-assessment is a reflective progress that allows learners to identify gaps in their knowledge and to consequently identify learning objectives (Alderson, 2005;

Little, 2005; Trofimovich et al., 2016). Reflective progress is a crucial feature of language learning because it increases learners' metacognitive awareness about their own progress and performances and promotes learner autonomy (Cossar-Burgess & Eberstein, 2013; Poehner, 2012). Unless learners understand how well they can perform, they cannot set appropriate learning targets and select suitable learning activities and materials. In addition, by knowing their own strengths and weaknesses, learners can make better decisions regarding how to apply their efforts and/or avoid mistakes (Carter & Dunning, 2008). Second, self-assessment is part of a learner-centered curriculum, as it leads to collaboration and shared responsibility between teachers and learners (Nunan, 1988). In a learner-centered curriculum, learners should be included from the process of evaluating their own learning achievement (Little, 2005). Third, learner self-assessment encourages learners to stay on task and use the target language; in such cases, time-on-task naturally increases. Finally, self-assessment potentially has other positive effects on language learning such as increases in motivation and confidence (Gardner, 2000).

Learner self-assessment can be a reliable measure of speaking proficiency because it provides "robust concurrent validity" as indicated by a meta-analytical study reported by Ross (1998, p. 16). Self-assessment is a valuable assessment tool because it correlates with teacher assessments (Malabonga, Kenyon, & Carpenter, 2005; Poehner, 2012). Ma and Winke (2019) found that self-assessment correlates with computerized oral proficiency interview measures where most students' gains in the computerized oral proficiency interview test were reflected in their own self-assessment gains. Learner self-assessment is therefore effective when it is integrated with other forms of measurements

(Little, 2005). On the other hand, many researchers have also argued that learner self-assessment are unreliable given that learner self-assessment and teacher assessment measures do not always correlate (Babaii, Taghaddomi, & Pashmforoosh, 2016; Brantmeier, Vanderplank, & Strube, 2012; Patri, 2002; Suzuki, 2009). Ross' meta-analytic study also indicated that some learners' assessment of their language proficiency predicts little variance in language performance. For example, learners tend to rate their own proficiency too low (Ma & Winke, 2019). Some researchers have even argued that learner self-assessment and objective measures of speaking performance correlate poorly if at all (Trofimovich et al., 2016). Peirce et al. (1993) found that learner self-assessments correlates weakly with teacher assessments, but they correlate higher with specific task benchmarks than with global benchmarks. Peirce et al. further found that correlations between self-assessment and objective teacher assessment are higher for lower-proficiency learners than higher-proficiency learners. Among the correlational analyses conducted, the highest correlations between self-assessed speaking proficiency were with specific task benchmarks ($r$ = .25) and the lowest were with global benchmarks ($r$ = .12).

However, several problems exist with learner self-assessment. First, learners' perceptions of their competence often differ from their true level of competence (Carter & Dunning, 2008). High performers can underestimate themselves relative to the perceptions of others, while low performers can inflate their scores by overestimating their ability (Ma & Winke, 2019; Trofimovich et al., 2016). Second, Carter and Dunning (2008) argued that accurate self-assessment is unlikely without training because learners do not have adequate information to properly conduct self-assessment. Third, learner

self-assessment can be culturally, socially, and regionally dependent. For example, Japanese EFL students have been reported to assess themselves more harshly than might be expected because of the cultural value placed on modesty (Matsuno, 2009).

## Communicative Language Teaching

Communicative language teaching (CLT) is a group of approaches to teaching speaking and a dynamic entity that has changed throughout the decades (Hughes, 2011). It is a humanistic and interactive method to second language teaching and learning and was predominant in the 1970s, 1980s, and 1990s (Savignon, 2005). A fundamental principle of CLT is that learners engage in communication based on their desire to communicate ideas concerning meaningful topics (Bygate, 1998). Communicative language teaching proponents have emphasized the importance of authentic communication as learners develop communicative competence from interacting with other individuals in communicative contexts. While there is no one universally recognized CLT model, it has been generally acknowledged that CLT classrooms include communicative purposes (Larsen-Freeman, 1986), the use of authentic materials (Li, 1998), and the negotiation of meaning in the target language (Sakui, 2004).

While CLT has become one of the most dominant methodologies for developing speaking skills, a major reoccurring criticism is that it over-emphasizes communication and meaningfulness, and therefore overlooks the explicit teaching of linguistic forms and syntactic accuracy (Savignon, 2005). This perception has arisen in part because some CLT proponents have discouraged direct grammar teaching because they believe that

explicit knowledge cannot be meaningfully transferred from teachers to students and language learning occurs during interactions in which learners use their linguistic resources to complete communicative tasks (Lochland, 2013). Such perceptions arose when pioneering research on the communicative competence of adult ESL learners indicated that learners who engaged in communication practice had the same level of grammatical accuracy as learners who engaged in traditional grammar-focused practice, while the communicative competence of the former group was better than the learners who had no communication practice (Savignon, 2005). However, despite the results of some studies, it is a misconception that grammar is never taught in CLT classrooms (Thompson, 1996). While authentic interaction is important, communicative competence is partially based on the use of accurate grammatical forms (Canale & Swain, 1980).

Communicative language teaching has exerted a strong influence in many countries, including EFL contexts. For example, CLT was mandated into the language education curriculum by seven Asian countries (Nunan, 2003). In Japan, CLT has also been widely practiced in English-language classrooms (Savignon, 2005; Takanashi, 2004). The Japanese Ministry of Education has implemented curricular reforms based on CLT—including the introduction of communicative syllabi, the implementation of the Japan Exchange and Teaching (JET) Program, and overseas in-service training for teachers (Savignon, 2005)—in an attempt to make classes more communicative. As a result, some English teachers in Japan have emphasized communication among students in the classroom without explicit grammar instruction (Savignon, 2005).

## Task-Based Language Teaching

Task-based language teaching has emerged as the dominant pedagogical approach in the communicative language teaching framework (Ellis, 2003, 2009b; Long, 2015; Nassaji & Fotos, 2011). TBLT has been diversely defined and this has resulted in different applications and implications (Ellis, 2003, 2018). Task-based learning was initially introduced to the field of language learning by Prabhu (1987), who argued that the focus of teaching needs to be placed on tasks, which he considered to be holistic units of communication. He also argued against the preselection of linguistic forms in part because he perceived language acquisition to be an implicit process that occurs when learners use language for communication. According to Prabhu, motiving and engaging tasks alone are enough for learners' linguistic development. Long (1985) and Long and Crookes (1992) emphasized the importance of language learning through communication but unlike Prabhu, Long advocated the need for learners to incidentally focus on form during interaction when communication breakdowns and misunderstandings occur. Long's (1991) definition of *focus on form* indicated that it occurs in primarily meaning centered discourse, observable (i.e., occurs interactionally), incidental, transitory, and extensive (i.e., several forms are attended to in a single lesson). Therefore, task-based language teaching is as an approach to course design in which learners learn by engaging in tasks that are meaningful, relevant, engaging, and stimulating, and which includes an incidental focus on form (Long, 1985). Long (2015) later defined task-based language teaching as "an approach to course design, implementation, and evaluation intended to meet the communicative needs of diverse groups of learners" (p. 5).

The concept of *task* has been diversely defined by different researchers (Butler, 2011; Ellis, 2003). From a broad non-pedagogical perspective, a task is "a piece of work undertaken for oneself or for others, freely or for some reward" and can be "the hundred and one things people do in everyday life, at work, at play and in between" (Long, 1985, p. 89). From a language learning perspective, Nunan (1989) defined task as "a piece of classroom work which involves learners in comprehending, manipulating producing or interacting in the target language while their attention is principally focused on meaning rather than on form" (p. 10). Skehan (1998) defined task to be "an activity in which meaning is primary, there is some communication problem to solve, there is some relationship to real-world activities, task completion has some priority, and the assessment of the task is in terms of outcome" (p. 95). Bygate, Skehan, and Swain (2001) defined task as "an activity which requires learners to use language, with emphasis on meaning, to attain an objective" (p. 11). Nunan, Skehan, and Bygate et al.'s definitions of task share the idea that a task is a language use activity in which there is a primary focus on meaning. These definitions of task can be compared with that of *exercise*, which is a language use activity that has a primary focus on linguistic form (Ellis, 2003). Ellis further defined tasks to be learner activity plans that involve primary focus on meaning, real-world language, the skills of listening, reading, speaking and writing, cognitive processes such as selecting, classifying, ordering, reasoning, and evaluating information, as well as clear communicative outcomes (pp. 9-10). In addition, Long (2015) defined task as the unit of analysis throughout the design, implementation, and evaluation of a

TBLT program, including ways student learning is assessed, by task-based, criterion-reference, and performance tests.

Richards and Rodgers (2001) argued that the claim that task-based language teaching is more effective than other language teaching approaches remains an ideology rather than fact. While there are many advantages of implementing task-based language teaching in the classroom, the approach is not without criticisms (Sheen, 1994; Skehan, 2003; Swan, 2005). One of the earliest criticisms of task-based language teaching relates to the need for teachers to be ready to provide unpredictable types of support when required by learners (Skehan, 2003). Another criticism is that learners do not learn the target linguistic forms as they are taught in traditional grammar-focused classrooms (Sheen, 1994) because task-based language teaching is often justified by the claim that linguistic regularities are acquired through noticing during communicative activities and should therefore be addressed incidentally by a focus on form implemented during a task.

However, the above criticisms are not entirely accurate because task-based language teaching has progressed from its original emphasis on incidental learning to the recognition of integrating a focus on linguistic form into meaningful tasks (Ellis, 2009; Long, 2015). Some SLA researchers influenced by cognitive psychology (e.g., DeKeyser, 2001; Robinson, 2001; Skehan, 1998) have argued that learners need opportunities to consciously notice their errors through a focus on form in order to improve grammatical accuracy. This idea supports Swain's Comprehensible Output Hypothesis (1985, 1995, 2005), which states that learners need to be pushed to produce precise and appropriate output so that they notice what they are missing when they produce the target language.

Unless learners notice linguistic form, they will be unlikely to stretch their interlanguage and address their linguistic deficiencies (Schmidt, 1990, 2001). Learners must be induced to pay conscious attention to linguistic forms in the input, especially forms they might otherwise ignore (e.g., third person *–s*) (Ellis, Basturkmen, & Loewen, 2001b). In sum, without noticing linguistic forms by focusing conscious attention on form, foreign language learners, including those with high communicative ability, are generally unable to develop a high level of grammatical accuracy (Yang & Lyster, 2010).

### *Focus on Form*

Within the context of task-based language teaching, the term *focus on form* (FonF) was first introduced by Long to describe a brief instructional focus on linguistic features during meaningful communication (Williams, 2005). Focus on form was later defined as the incidental attention that teachers and learners pay to form in the context of meaning-focused instruction (Ellis et al, 2001b). Therefore, the focus on form approach places a primary focus on the communication of meaning and any focus on linguistic features occurs incidentally when it is triggered by problems that occur during communication. *Focus* refers to the act of leading learners' attention to features of the target language during meaningful communication; that act is typically brief and it occurs during the processing of meaning for communication (Williams, 2005). *Form* refers to structural linguistic features that are noticed during communication. Targeted forms need to originate from problems experienced by learners in communicative tasks (Ellis, Basturkmen, & Loewen, 2001a). The two general approaches to *focus on form* are the

intensive and extensive approaches (Rahimpour, Salimi, & Farrokhi, 2012). The intensive approach involves repetitive exposure to a single preselected linguistic feature, whereas the extensive approach refers to nonrepetitive exposure to numerous linguistic features in a single lesson. The extensive approach involves instructors responding to all linguistic errors that occur incidentally during communication or when communicative breakdowns occur. Because the intensive approach is limited to one or a few target forms pre-selected by teachers based on a prior analysis of linguistic problems experienced by the learners, it can also be considered preemptive, given that teachers anticipate linguistic problems before they occur (Ellis et al., 2001b).

**Form-Focused Instruction**

Ellis (2001) used the umbrella term *form-focused instruction* (FFI) to refer to both incidental and planned instruction that directs learner attention to target language linguistic forms. The definition of form-focused instruction varies, as it can be both implicit and explicit, and can range from pedagogical activities such as planning and repetition, to teacher modelling, teacher explanations of target forms, and teacher feedback (Norris & Ortega, 2000). Deeper processing of target forms can be encouraged through form-focused instruction where form can assist interlanguage development (Ellis et al., 2001b). Form-focused instruction can encourage learners to attend to forms in the input that they have not yet acquired (Williams, 2005) and compare their current interlanguage rules and the target language rules in terms of both input and output processing (Doughty, 2001). In other words, effective FFI leads learners to compare

forms in the input to representations stored in their long-term memory or traces left in short-term memory (Doughty, 2001). However, direct comparisons are unlikely because input must be processed before it can be stored and comparisons made. Therefore, if target language input cannot be stored in long-term memory, then the noticing-the-gap activities that come after storing input would have limited effectiveness (Williams, 2005). It is therefore necessary to implement approaches that encourage target language input to be stored and incorporated into the developing interlanguage system. Izumi (2002) conducted an experimental study to compare the effects of output and enhanced input on noticing and development. He reported that participants with enhanced input exhibited more noticing but they did not improve more in the target language.

**The Effects of Planning on CAF**

Many FFI studies have been centered on the effects of planning (Ahmadian & Tavakoli, 2011; Mochizuki & Ortega, 2008; Sangarun, 2005; Yuan & Ellis, 2003) because planning can lead to improved working memory functioning and a better conceptualization of the task (Ellis, 2009a). Planning also improves the formulation of ideas and the articulation of phonemes because learners are likely to have accessed linguistic knowledge while planning and as a result, they find it relatively easy to access the same information when performing the task. Planning can be divided into two major types: pre-task planning and within-task planning (i.e., online planning) (Ahmadian & Tavakoli, 2011; Ellis, 2005a, 2009a; Yuan & Ellis, 2003).

Pre-task planning can be further divided into strategic planning and rehearsal (repetition). Strategic planning involves learners considering what they are going to say in terms of content and language prior to performing the task whereas rehearsal refers to the opportunity of doing the task once as preparation (can be partial or the whole task) before actually performing the task. A large of researchers have investigated the effects of pre-task planning on oral proficiency development by manipulating different lengths of planning time using narrative story-telling tasks (Elder & Iwashita, 2005; Kawauchi, 2005; Mochizuki & Ortega, 2008; Tavakoli & Skehan, 2005). It was found that pre-task planning was effective at improving oral fluency in most studies (e.g., Ellis, 2005b; Foster & Skehan, 1996; Mehnert, 1998; Ortega, 1999), effective at improving syntactic complexity in some studies (e.g., Skehan, 2009a), but almost never improved syntactic accuracy (two exceptions were Mochizuki & Ortega [2008] and Sangarun [2005]).

Researchers (e.g., Bygate, 1996, 2001; Baleghizadeh & Shahri, 2017) have reported that rehearsal benefits subsequent performances of the same task. Rehearsal also benefits oral fluency and syntactic complexity development (Bygate, 1996, 2001), but it exerts almost no positive effect on syntactic accuracy (Bygate, 2001). In addition, the effects of rehearsal become more significant when it is used in conjunction with corrective feedback; this combination led to improvements in complexity, accuracy, and fluency in a study by Sheppard (2006). However, it was found that the effects of rehearsal planning on one task are not transferable to other tasks even when equivalent tasks are performed (Bygate, 1996, 2001). With within-task planning, research has

indicated that careful planning during the task leads to greater complexity and accuracy in L2 learners' oral production (Ahmadian & Tavakoli, 2011; Yuan & Ellis, 2003).

### The Effects of Task Repetition on CAF

In addition to the effects of planning, a large number of form-focused instruction researchers have investigated the effects of task repetition on speaking (e.g., Bygate, 1996, 1999, 2001; Bygate & Samuda, 2005; Lambert, Kormos, & Minn, 2017). Some researchers have treated task repetition as one type of pre-task planning in which learners repeat the same or a slightly altered task partially or wholly (Bygate & Samuda, 2005). According to Levelt's (1989) speech production model, repetition reduces learners' processing load for conceptualization, which frees up working memory capacity. Fukuta (2016) further explained the mechanism behind the effects of repetition on speaking development by noting that learners are primarily concerned with content rather than linguistic form when engaging in a task the first time, and this focus allows them to familiarize themselves with the content of the task, which allows them to free up working memory capacity and focus attention on linguistic form.

Task repetition also improves speaking fluency because it leads to more automatized speech (Ahmadian, 2011; Bygate, 1996, 2001; Bygate & Samuda, 2005; Ellis, 2005b). It can also improve formulation and articulation because learners access much of the same linguistic and procedural knowledge they accessed initially when they repeat a task; this process makes formulation and articulation in subsequent performances easier and frees up working memory capacity (Bygate, 2001). Evidence for positive

effects on working memory has been provided by studies in which repetition has been found to improve syntactic complexity (Ahmadian, 2011; Bygate, 2001; Gass et al., 1999; Skehan, 2009a), which presumably occurs because learners can recycle some of the cognitive work done in the area of morphosyntax in subsequent performances. This increased memory capacity makes it possible for learners to focus on new and more complex syntactic structures.

Past studies have found that task repetition helps speaking fluency in the short term (e.g., Ahmadian, 2011; Bygate, 1996, 2001; Bygate & Samuda, 2005; Ellis, 2005b). Most studies on task repetition have been cross-sectional (see Kanda, 2015 for an exception), so the researchers were unable to confirm the effects of longitudinal task repetition on speaking. Lambert et al. (2017) conducted a study examining the effects of task repetition as a form-focused instruction intervention on speaking fluency over six performances in a single session. In the first speaking task, the largest speech rate gains occurred in the first three repetitions and the gains continued until the fifth repetition. In contrast, the number of self-repairs started to decrease from the fifth repetition; this development indicated increased fluency. However, for the second narration task, the number of self-repairs decreased from the third performance, but they increased in the sixth performance. Despite finding significant gains in speaking fluency over six repetitions, the performances occurred in one 90-minute session; thus, the effects of task repetition on speaking proficiency longitudinally over time is still unknown.

In contrast to the positive effects on oral fluency (e.g., Lynch & Maclean, 2000), most previous studies on the effects of task repetition have found small gains in speaking

accuracy (Fukuta, 2016; Gass et al., 1999; Kanda, 2015; Sangarun, 2005). The insignificant accuracy gains are probably the result of learners primarily focusing their attention on formulating and processing content and secondarily on linguistic form (i.e., syntactic complexity). Finally, although repetition benefits subsequent performances in some ways, it is still unclear whether these beneficial effects transfer to different tasks even when the task is similar to the original task (Bygate, 1996, 2001).

### The Effects of Explicit Form-Focused Instruction on CAF

Numerous TBLT researchers have recognized the effectiveness of integrating form-focused instruction into meaningful communicative contexts (Ellis, 2013, 2018; Long, 2015; Nassaji & Fotos, 2011; Skehan, 2018). Many studies have found that explicit attention to form in communicative contexts is effective (Spada, 1997; Norris & Ortega, 2000) especially in foreign language settings where the target language input is limited (Fotos, 2002). Many previous researchers have used pre-task planning and repetition to investigate the effects of form-focused instruction by directing learners' attention to linguistic form. However, as noted in most of the pre-task planning and task repetition studies reviewed above, they failed to show syntactic accuracy gains (Williams, 2005). This lack of syntactic development suggests that there might have been a lack of explicitness in previous pre-task planning and/or task repetition interventions. Thus, there is a need for more explicit form-focused approaches that explicitly draw learners' attention to linguistic form.

Explicit form-focused instruction directs students' attention to specific language learning objectives through demonstrations, explanations, and focused practice (Doughty & Williams, 1998). Explicit form-focused instruction is considered to be effective for adult foreign language learning, especially for EFL learners, because it reduces ambiguity and facilitates cognitive mapping between form, meaning, and the use of linguistic expressions (DeKeyser, 1998; Norris & Ortega, 2000; Spada & Tomita, 2010). Through explicit form-focused instruction, teachers can draw learners' attention to specific morphosyntactic forms that can enhance their ability to speak accurately (Goh & Burns, 2012). In addition, explicit instruction can help learners notice inappropriate forms in their own output. As mentioned above, unless learners notice errors in their own utterances, it is unlikely that they will modify their spoken language and achieve greater accuracy (Swain, 1985).

Noticing, however, is a complex process involving both meaning and form. Learners require time to gain intake and internalize new forms into their interlanguage (Batstone, 1996). Therefore, teachers need to provide recurring opportunities for learners to notice linguistic forms until they can internalize the knowledge. Explicit form-focused instruction directs learners' attention to target forms by raising the learners' consciousness of the target form(s). How explicit FFI encourages the noticing of particular forms remains an ongoing question in task-based language teaching. In the present study, explicit form-focused instruction refers to grammar instruction partly through the delivery of metalinguistic explanations by the classroom teacher.

**Metalinguistic explanations.** While early views against direct grammar instruction stated that grammar can only be acquired unconsciously through exposure to communicative input (Krashen, 1981), most researchers have concluded that grammar instruction helps learners acquire grammatical features more accurately and quickly (e.g., Ellis, 2005b; Han, 2004). However, the effect of grammar instruction for developing speaking skills is difficult to measure, as those effects cannot be captured in oral output as clearly as in written, discrete-point grammar tests (Ellis, 2005b). Although grammar instruction does not alter the natural order of acquisition, it enables learners to progress more rapidly along the natural route (Ellis, 2005b). Therefore, the controversy over grammar instruction is no longer on whether form-focused instruction is effective or not, but rather on what type of instruction should be used (Spada & Lightbown, 2008).

Ellis (2005b) has promoted three commonly practiced types of grammar instruction—explicit instruction, production practice, and negative feedback—and suggested that they can be combined in various ways. Explicit instruction can be divided into didactic and discovery instruction. Didactic instruction refers to direct teacher explanations of grammar, while discovery instruction uses various types of consciousness-raising tasks that require learners to work out for themselves how the morphosyntactic forms work. Production practice refers to instruction that requires learners to produce output containing the target form(s) through both controlled and functional tasks. Controlled instruction gives learners guidance in producing output containing the target forms (e.g., by filling in blanks), whereas functional instruction requires learners to produce original sentences containing the target forms in specified

contexts. Negative feedback can be delivered implicitly through recasts or explicitly through prompts in response to students' language output. Implicit feedback models the correct target forms without explicitly indicating the errors, whereas explicit feedback clearly indicates the presence of an error.

Muranoi (2000) examined the acquisition of English articles from interaction enhancement, which is a form-focused instruction treatment where interactional modifications are provided for learners where they are led to produce modified output in problem-solving tasks. The participants were 91 first-year Japanese students in three intact classes. Two treatment groups participated in the study: interaction enhancement plus formal debriefing, interaction enhancement plus meaning-focused debriefing, and a control group. Between the pretest and posttest, three 30-minute training sessions were provided for the experimental group participants during normal class hours. The results indicated that interaction enhancement was an effective type of form-focused instruction that positively affected English learning, especially when combined with grammar instruction. Muranoi (2000) argued that form-focused instruction that systematically directs learners to target forms is necessary in communicative language teaching classrooms. Muranoi's study produced significant results regarding the effect of form-focused instruction on syntactic accuracy. However, it was not longitudinal as the researcher provided only three interventions in a one-week period. Therefore, to enhance the effectiveness of explicit form-focused instruction on accuracy development, longer interventions combined with more explicit forms of instruction are desirable.

Mochizuki and Ortega (2008) examined the effects of form-focused instruction in terms of pre-task planning and grammatical guidance on beginner-level learners. First-year Japanese high school students were asked to do a story retelling task with a partner. They were divided into three groups: no planning, five-minute planning, and five-minute planning with guidance using a grammar handout on English relative clauses. Students stories were transcribed and narrated for global complexity, fluency, and accuracy. The study found that guided planning group produced more and more accurate relative clauses in their narratives than the other two groups, while at the same time produced similar levels of complexity and fluency with the other two groups. Mochizuki and Ortega's (2008) study showed that guided planning with grammar focus can be integrated into communication based classrooms.

While almost no previous researchers have examined the effects of form-focused instruction on speaking proficiency development longitudinally, Kanda (2015) conducted a study examining the long-term effects of form-focused instruction on speaking proficiency development by examining the changes in 15 Japanese high school students over one academic semester. Similar to most previous studies on form-focused instruction (e.g., Ahmadian & Tavakoli, 2011; Lambert, Kormos, & Minn, 2017; Mochizuki & Ortega, 2008; Sangarun, 2005), Kanda used task repetition and planning as interventions. Kanda found that her low-proficiency participants improved in terms of speaking proficiency, oral fluency, syntactic complexity, and syntactic accuracy through pre-task planning, on-line planning, and task repetition during the academic year. Kanda suggested that the combined effects of pre-task planning, on-line planning, and task

repetition can effectively improve oral fluency, syntactic complexity, lexical complexity, and syntactic accuracy in task-based teaching classrooms because they drive learners' attention to form-meaning mappings. However, the higher-proficiency learners only improved in oral fluency and lexical complexity; they did not improve syntactic accuracy. In addition, Kanda found that oral fluency, syntactic complexity, and syntactic accuracy (i.e., the ratio of accurate past tense usage) generally increased from Sessions 4 and Session 8 whereas syntactic complexity increased significantly throughout the academic semester. However, Kanda did not find gains in syntactic accuracy in the initial weeks of the research period. Kanda commented on the lack of explicit instruction of linguistic form in her study and suggested that if explicit instruction had been provided about how and when to use past tense forms during the pre-planning time, the learners might have achieved more accurate use of past tense forms.

**Corrective feedback.** Corrective feedback is one type of form-focused instruction because it directs learners' attention to their linguistic errors. Mackey (2006) found that feedback was effective at promoting noticing of L2 forms. Numerous studies on the effects of corrective feedback have produced robust evidence for its effectiveness (e.g., Li, 2010; Lyster & Saito, 2010), while a lack of negative feedback can inhibit the development of syntactic accuracy (Goh & Burns, 2012). According to Swain's Comprehensible Output Hypothesis, learners need to be pushed to notice discrepancies between their language production and target language forms; the more learners notice non-target linguistic forms in their output, the higher the probability they will improve on

the target forms. Corrective feedback and other types of form-focused instruction can be combined to create multiplier effects (Lyster, 2004) that greatly enhance the probability that target linguistic forms will be made salient. Notwithstanding these positive findings, there is a continuing controversy over the effectiveness of corrective feedback because there is evidence that implicit types of feedback, such as recasts, are not always interpreted by learners as error correction. If learners do not realize that an utterance is corrective, they are unlikely to notice the gap between their production and the target language rule (Williams, 2005). In addition to the concerns with using implicit types of feedback, feedback can also be ineffective itself when extensive correction is given, in which learners receive feedback on too many types of errors (Williams, 2005). In addition, even if learners notice errors in their own output, it is not certain that they will successfully integrate what they have noticed into their interlanguage (Nicholas, Lightbown, & Spada, 2008). Lyster (2004) further argued that only explicit forms of feedback such as prompts can push learners to precisely and appropriately reformulate their output; he therefore doubts the value of implicit feedback such as recasts as a classroom technique. However, there are some criticism against corrective feedback. For instance, teacher feedback can be ineffective due to the teachers' use of vague and "rubber stamp" comments as well as an overreliance on grammar correction (Paulus, 1999). In addition, ineffective teacher intervention can result in students' inattention and negative attitudes toward receiving feedback (Paulus, 1999).

**Peer Corrective Feedback**. Besides teacher feedback, research has indicated that peer feedback can also improve the process and product of language learning (Lee & Tajino, 2008). Although it is used more commonly for improving the quality of writing tasks (Yang, Badger & Yu, 2006), peer feedback is becoming increasingly common in speaking tasks and more prevalent in foreign language classrooms because it resolves the time and class size constraints that make it difficult for teachers to give individualized feedback (Yang, Bedger, & Yu, 2006). In addition, while higher-achieving students seem to respond positively and benefit from teacher feedback, some lower-achieving students respond negatively to it, as they cannot comprehend the comments given by their teachers (Guénette, 2007). Peer corrective feedback, on the other hand, has been shown to have many positive effects. While research has indicated that teachers tend to make more comments for improving grammatical features, learners can generate more comments during peer review for improving the content, organization, and vocabulary (Paulus, 1999). In addition, besides beneficial effects on language performance, peer feedback also has advantages such as developing critical thinking, learner autonomy, and social interaction among students (Yang, Badger, & Yu, 2006). Positive social interaction is important because it can decrease learner anxiety as well as increase their desire to use the target language.

Therefore, some implicit forms of peer corrective feedback have been criticized for their quality as they generate unclear comments, which becomes a drawback to the abovementioned critique of the effectiveness of corrective feedback. While explicit peer feedback can potentially overcome the lack of clarity of implicit feedback, it is often

doubtful that language learners can give quality explicit comments because of their limited knowledge, experience, and relatively low language proficiency (Saito & Fujita, 2004). Although some learners' lack the ability to give appropriate comments on erroneous morphosyntactic forms, peer feedback can be useful where content, organization, and vocabulary are concerned (Paulus, 1999). As the result, learners are unable to effectively give and receive peer corrective feedback.

## Learner Affect Toward Speaking English

Affective variables are an important aspect of second language learning and teaching as they partly determine learners' experiences and behaviors (Dörnyei, 2005; Incecay & Dollar, 2011; Loewen, Li, Fei, Thompson, Nakatsukasa, Ahn, & Chen, 2009). Although how learners feel about the target language does not always reflect their performances (Incecay & Dollar, 2011), past researchers have pointed out that affective states can predict learner behaviors, influence achievement, lead to higher proficiency (Mori, 1999; Yang, 1999; Yashima, 2002), and increase participation in classroom activities (Borg, 2003). Therefore, understanding learner affect is important for understanding learners' cognitive processes and language learning behaviors (Swain, 2013). In the following sections, I review literature concerning three affective variables: classroom English speaking anxiety, English speaking self-efficacy, and desire to speak English. While no other researchers have investigated the effects of form-focused instruction on learner affect in task-based classroom, Van Batenburg et al. (2019) found that form-focused instruction decreases learners' willingness to communicate (WTC),

increases their self-confidence when combined with interaction strategy instruction, and does not influence their enjoyment level.

**Classroom English Speaking Anxiety**

Anxiety has been one of the most widely researched affective variables in the field of second language learning and teaching (Horwitz, 2010; Horwitz, Horwitz, & Cope, 1986). Understanding learner anxiety is important because researchers have generally reported that more proficient classroom learners are generally less anxious than less proficient learners (Liu, 2006). Language learning anxiety is a subjective feeling of tension, apprehension, nervousness, and concern related to language learning (Horwitz et al., 1986). As such, anxious learners perceive speaking activities to be distressing (Müge Satar & Özdener, 2008). Anxiety can interfere with language learning and performance because it can interfere with learners' comprehension of linguistic input, production of output, and retention of newly acquired language (Horwitz et al., 1986; MacIntyre & Gardner, 1991). On the other hand, anxiety can also have positive effects on language learning because more anxious learners can set higher personal performance goals and be more concerned about errors (Ewald, 2007; Gregersen & Horwitz, 2002; Kitano, 2001).

Anxiety toward the language classroom is situation-specific. Horwitz et al. (1986) investigated situation-specific anxiety and developed the Foreign Language Classroom Anxiety Scale for measuring foreign language learning anxiety in the classroom. Unlike previous studies focused on general language anxiety, Horwitz et al. argued that anxiety specific to foreign language classroom learning needs to be defined and distinguished

from general second language anxiety because foreign language anxiety is a distinct combination of "self-perceptions, beliefs, feelings, and behaviors related to classroom language learning arising from the uniqueness of the foreign language learning" (p. 128). Speaking is considered the most anxiety-provoking element in foreign language classrooms (Cheng, Horwitz, & Schallert, 1999). Most EFL learners have a high level of speaking anxiety compared to reading, listening, and writing because speaking is the skill to which the students' language ego, the identity individuals develop in reference to the language they speak, is most vulnerable (Gkonou, 2014). Speaking anxiety is considered to be a determinant of speaking performance (Mak, 2011; Sparks & Ganschow, 1995). Teacher instruction can increase speaking anxiety (Saito, 2004; Trang, Baldauf Jr., & Moni, 2013) as some learners, for example advanced-level learners, have more fear of being negatively evaluated by their teachers during instruction (Kitano, 2001).

**English Speaking Self-Efficacy**

Self-efficacy was introduced by Bandura (1977) to describe learners' personal judgments of their own capabilities to attain designated goals. SLA researchers have described learners' self-perceptions of their own performances using terms such as *self-confidence* (Gardner, Tremblay, & Masgoret, 1997), *linguistic self-confidence* (Clément, Dörnyei, & Noels, 1994), *self-perceptions*, and *self-ratings* (MacIntyre, Noels, & Clément, 1997). However, unlike these terms, self-efficacy is task specific and specific to the context of the situation (Taipjutorus, Hansen, & Brown, 2012). Self-efficacy beliefs are important because self-efficacious learners feel competent in their ability to

accomplish challenging tasks and from the positive experiences of successful task completion, such students become more willing to engage in more difficult tasks in the future (Mills, 2014). In addition, self-efficacy has direct and indirect positive and negative effects on different aspects of language learning (Piniel & Csizér, 2013), such as directly determining language learning success (Mills, Pajares, & Herron, 2006), and indirectly increasing learner motivation, engagement, persistence, and effort, as well as decreasing learner anxiety (Bandura, 1988). Therefore, Bandura (1997) also argued that self-efficacy is a more reliable predictor of learner behavior and achievement than other affective factors because in his view, it is the most influential arbiter of human agency and influences decision making.

According to Bandura (1997), there are four sources of self-efficacy: performance accomplishments, vicarious experience, social persuasion, and physiological and emotional states. Learners have higher self-efficacy if they can relate a current task to similar experiences they have successfully completed previously, if they can observe peers successfully completing a similar task, and if they have little anxiety toward a task (Bandura, 1997). These sources of self-efficacy can form a cyclic relationship in which positive past experiences lead to higher self-efficacy, which leads to greater effort and more positive experiences, which subsequently lead to higher levels of self-efficacy.

While a number of self-efficacy studies have already been conducted in the field of second language learning and teaching, few studies have concerned speaking (two exceptions are Apple, 2011 and Leeming, 2017). Leeming conducted a longitudinal study measuring changes in English-speaking self-efficacy with Japanese university students

over one academic year using a speaking self-efficacy questionnaire and student interviews. Leeming constructed nine questionnaire items to measure his participants' perceptions of their own capabilities toward specific speaking-related tasks. He found significant growth in speaking self-efficacy over the academic year. Apple, on the other hand, investigated *perceived foreign language speaking self-competence* to examine learners' perceptions of their abilities to perform specific classroom-related speaking tasks. Apple reported that learners' self-perceived English-speaking competence is influenced by personality characteristics such as extraversion/introversion, agreeableness, emotional stability, conscientiousness, and intellect/imagination as well as by social situations such as past English classroom experiences and current English classroom perception. Although Apple did not use the term *speaking self-efficacy*, Table 1 indicates that many of his questionnaire items overlapped with Leeming's items as they all concerned the ability to perform speaking tasks. All of Leeming's nine questionnaire items concerned a communicative English classroom (Taipjutorus et al., 2012). On the other hand, Apple's questionnaire items 1 to 6 were specifically related to a communicative English classroom while items 7 and 8 concerned general speaking ability outside the classroom. However, while items 7 and 8 were related to general speaking ability, they also concerned real-life situations that are often incorporated into communicative classrooms.

Table 1. *English Speaking Self-Efficacy Items from Leeming (2017) and Apple (2011)*

| | Leeming (2017) | | Apple (2011) |
|---|---|---|---|
| 1 | I can give my opinion in English | 1 | I can give an individual speech in English in front of my class |
| 2 | I can speak English fluently when giving a presentation in front of the class | 2 | I can introduce myself in English during pair work to a classmate |
| 3 | I can enjoy conversation in English | 3 | I can talk about my hobbies in English during pair work with a classmate |
| 4 | I can speak English fluently when taking part in a group discussion | 4 | I can give an English speech in a group of my classmates |
| 5 | I can use my study skills well in English Communication class | 5 | I can give an English presentation as part of a group in front of my class |
| 6 | I can learn the material for English Communication class | 6 | I can easily join in a conversation in a group of native English speakers |
| 7 | I can do very well in English Communication class | 7 | I can give street directions in English to a foreigner |
| 8 | I can receive a good grade in English Communication class | 8 | I can order food in English in a restaurant |
| 9 | I can understand what is taught in English Communication class | | |

**Desire to Speak English**

Desire to speak English concerns learners' positive affective disposition for speaking English, integrating into English speaking culture, and identifying with members of the target-language community (Gardner, 1985; Yashima, 2002; Yashima Zenuk-Nishide, & Shimizu, 2004). While closely related to *Integrativeness* (Gardner, 1985) and *International posture* (Yashima, 2002; Yashima et al., 2004), desire to speak English differs from these two constructs in that it captures the tendency of learners to relate themselves to the English-speaking community rather than the entire international community (Apple, 2011). The desire to use the target language might be one of the most

well-known affective variables related to willingness for communication (MacIntyre, 2002). Although the relationship between desire to speak the target language and the oral performance is controversial, it has been found that the desire to speak the target language influences learners' task engagement (MacIntyre, Babin, & Clément, 1999). While the desire to speak the target language does not determine performance outcomes, higher levels of willingness to use the target language increase the opportunity for language practice and learning, which likely facilitates the language learning process (MacIntyre, Baker, Clément, & Conrod, 2001).

Desire to speak English positively affects learners' motivation, which leads to higher proficiency and greater self-confidence, which, in turn, partially accounts for learners' willingness to communicate in English. Desire to speak English correlates strongly with attitudes toward the learning situation, motivation, integrative orientation, and instructional orientation (Masgoret & Gardner, 2003). In L1 speech production, the desire to speak is considered a personality trait (MacIntyre, Dörnyei, Clément, & Noels, 1998); however, the desire to speak an L2 is likely to change depending on factors such as familiarity with the interlocutor, the formality of the situation, and past experience using the target language.

Two factors are closely related with learners' desire to speak the target language: communication apprehension and self-perceived competence (Van Batenburg et al., 2019). Communication apprehension refers to learners' anxiety toward using the target language (Horwitz & Young, 1991), whereas perceived competence refers to the learners' self-evaluation of their ability to speak the L2. The perception of ability to

speak the target language might be more important than actual target language competence when determining the desire to speak the target language (MacIntyre, Baker, Clément, & Donovan, 2003). Besides communication apprehension and self-perceived competence, other affective factors such as introversion, reticence, and shyness influence learners' desire to speak English (Yashima et al., 2004).

Besides affective factors, the style of instruction can also affect learners' desire to speak the target language; however, it is unknown what effects form-focused instruction has on learners' desire to speak the target language, as no previous studies are available (see Van Batenburg et al., 2019 for an exception). Nevertheless, despite the lack of previous investigations into the effects of form-focused instruction on learners' willingness to speak the target language, some effects are expected because different learners have different preferred instruction styles (Trofimovich, Lightbown, & Halter, 2013). Dewaele, and MacIntyre (2014) found that learners prefer unpredictable tasks over form-focused tasks, as the former give them a sense of autonomy. Van Batenburg et al. (2019) also reported that learners have decreased the desire to speak the target language from receiving form-focused instruction. However, the effects of form-focused instruction are not all negative, as it was found in the same study that learners' self-confidence increased because structured tasks allowed them to achieve the task goals in a structured environment with ease and in relative safety. Despite the unclear results surrounding the effects of form-focused instruction on desire to speak the target language, it was suggested that lower-proficiency learners might be positively influenced from explicit form-focused instruction interventions (Lam, 2006).

## Gaps in the Literature

The first gap in the literature is that few TBLT speaking development studies have produced significant gains in syntactic accuracy (see Yuan & Ellis, 2003 for an exception). Previous investigations of the effects of form-focused instruction have predominantly used pre-task planning and task repetition as interventions. Many previous researchers have shown that these two task manipulations usually affect oral fluency and syntactic complexity; their influence on syntactic accuracy development is typically weak or non-existent (Ahmadian & Tavakoli, 2011; Bygate & Samuda, 2005; Sangarun, 2005; Yuan & Ellis, 2003).

The second gap concerns the small number of studies concerning the effects of grammar instruction on speaking development in task-based contexts. Past researchers have investigated the effects of form-focused instruction on speaking development, but they primarily used planning and task repetition as interventions (e.g., Bygate & Samuda, 2005; Lambert et al., 2017; Mochizuki & Ortega, 2008). Bygate and Samuda looked at the effects of planning and task repetition and found that Mochizuki and Ortega looked at the effects of different types of planning on English relative clauses and found that guided planning allowed participants to produce a larger number of relative clauses and more accurate relative clauses in oral narratives. Yang and Lyster (2010) defined corrective feedback as form-focused instruction and looked at the effects of different levels of form-focused instruction on the accurate production of regular and irregular past tense verbs. Ellis, Loewen, and Erlam (2006) also looked at the effects of feedback on regular past tense *-ed* and comparative *-er* and found that explicit feedback was more

effective than implicit feedback. Although these studies are informative, they show that past researchers have rarely looked at the effects of explicit teacher-led grammar instruction on speaking in TBLT context.

The third gap is that few researchers investigating the effects of form-focused instruction have used a longitudinal intervention. For instance, Mochizuki and Ortega (2008) examined the effects of guided planning on speaking production by using a one-time intervention. Lambert et al. (2017) examined the effects of six performances of task repetition on speaking production in one session. Xu and Lyster (2014) also examined the effects of form-focused instruction on speaking production in one session using explicit grammar instruction. While the above studies have brought great value to the TBLT field, the researchers were unable to examine learners' acquisitional trajectories using cross-sectional designs; only longitudinal designs can show the learning trajectories taken by individual learners.

The fourth gap concerns the lack of understanding of the effects of form-focused instruction on learner affect including speaking anxiety in the English classroom, speaking self-efficacy, and desire to speak English (Van Batenburg et al., 2019). Although past researchers have investigated the effects of form-focused instruction on language learning, these studies were mostly conducted from a cognitive perspective, as the researchers focused on changes in syntactic complexity, syntactic accuracy, and oral fluency. Almost no studies were conducted on the effects of pedagogical instruction on learner affect (see Van Batenburg et al., 2019 for an exception). Grammar instruction not only influences the development of language skills, it also provides psychological

benefits because it reassures learners that they are receiving instruction and therefore feel that the delivery of knowledge is taking place (Graus & Coppen, 2016). Some researchers have indicated that learners, especially low-proficiency learners with relatively little learning experience, prefer form-focused, explicit, deductive instruction (Graus & Coppen, 2016). However, other researchers have indicated that learners generally express unfavorable attitudes toward grammar instruction (Loewen et al., 2009).

Teachers' instruction style is known to influence learners' classroom anxiety (Dolean, 2016; Kitano, 2001; Young, 1991); therefore, while other factors such as assessment methods, pedagogical approaches, student-teacher relationships, and curriculum structure (Trang et al., 2013) exert an influence, researchers have identified teacher instruction to be an important determinant of learner affect (Yan & Horwitz, 2008; Young, 1991). A current gap in our understanding concerns the psychological effects of form-focused instruction on learners speaking anxiety, speaking self-efficacy, and desire to speak English.

The fifth gap concerns the lack of understanding of learner self-assessment, especially in speaking performances. While many researchers have investigated learner self-assessment, they have produced varied results; some researchers have found learner self-assessment to be an effective measure of language proficiency (e.g., Little, 2005) while most others have not (Babaii et al., 2016; Brantmeier et al., 2012; Patri, 2002; Suzuki, 2009). In addition, not enough studies have been conducted on learners' self-assessment of speaking proficiency. As indicated by previous literature, learner self-assessment is an important tool for learners to better understand themselves. Therefore,

there is a need to continue investigating how learners would assess their language proficiency, especially speaking.

## Purposes of Study

The first purpose was to examine changes in syntactic accuracy, an issue that has not been clearly clarified by previous speaking development studies in the TBLT field. It is possible that previous TBLT researchers were unable to produce significant syntactic accuracy gains because they did not use explicit form-focused instruction. In contrast, I examined the changes to global syntactic accuracy and the use of three past verb tenses—simple past, past progressive, and past perfect. These three past tense forms were explored because the target English-language learners, Japanese university students, make frequent errors with past tense forms when speaking English.

The second purpose of this study was to examine the effects of grammar instruction on speaking development in a TBLT context by examining changes in learners' syntactic complexity, syntactic accuracy, and oral fluency. Explicit form-focused instruction can reduce the ambiguity of language learning by directing students' attention to specific learning objectives in a structured environment through demonstrations, explanations, and their completion of activities in which they practice target linguistic forms. Besides Ellis' (2006, 2009b, 2013) statements concerning the importance of integrating form-focused instruction into task-based learning, the addition of explicit form-focused instruction to TBLT studies is also supported by Norris and Ortega's (2000) meta-analysis study that indicated that explicit form-focused instruction

is more effective than implicit form-focused instruction. In this study, I investigate the influence of form-focused instruction on speaking development by comparing the speaking performance of participants from two treatment groups and a comparison group that received no form-focused instruction. The first treatment group received weekly grammar instruction from the classroom teacher, while the second treatment group received weekly grammar instruction plus peer feedback.

The third purpose was to examine oral proficiency development longitudinally. Due to the complex nature of collecting, transcribing, analyzing, and interpreting spoken data, most past studies concerning speaking proficiency development were cross-sectional. Therefore, my aim was to clarify the relationships and changes among syntactic complexity, syntactic accuracy, and oral fluency after exposing the participants to a seven-week intervention. A longitudinal design was used to try to describe the dynamic and non-linear nature of second language development.

The fourth purpose was to further examine the effect of form-focused instruction on speaking proficiency development in terms of three affective variables—speaking anxiety, speaking self-efficacy, and desire to speak English. An examination of past studies on foreign language anxiety by Horwitz (2010) indicated that early researchers focused on the effects of anxiety on performance whereas later researchers were more concerned with the sources of foreign language anxiety, its changes, its relationship with other learner factors, anxiety toward specific aspects of language learning, and instructional strategies. Regardless of the different foci adopted in earlier studies, they did not look at the effects of form-focused instruction at changing learner affective variables.

The fifth and final purpose of this study was to investigate the effects of form-focused instruction on speaking as assessed by learners' self-assessment. While numerous researchers have examined the development of learners' speaking proficiency in TBLT contexts (e.g. Ellis, 1995, 2001, 2003, 2005a, 2005b, 2009a, 2009b, 2018; Mochizuki & Ortega, 2008; Skehan, 2009a, 2009b, 2014, 2018), none used learner-centered measurements. Learners' assessment of their own language performance is important because it increases their awareness, motivation, investment, and agency, which is their investment in their own learning (Gardner, 2000; Ma & Winke, 2019). It has pedagogical and practical values because it not only helps classroom teachers to better understand learner mentality but also offers the opportunity for teachers to reflect on their own assessment (Alderson, 2005; Trofimovich et al., 2016). Learner self-assessment offers an alternative perspective for understanding the development of second language speaking proficiency (Alderson & Banerjee, 2001; Brown & Hudson, 1998).

## Research Hypotheses

The present study is guided by the following five research hypotheses:

Hypothesis 1a. Participants who receive form-focused instruction (FFI group) and participants who receive form-focused instruction and peer feedback (FFI + PF group) will improve syntactic accuracy because form-focused instruction explicitly directs them to pay attention to the use of syntactic forms (DeKeyser, 1998; Goh & Burns, 2012; Norris & Ortega, 2000; Spada & Tomita, 2010). Participants in the intervention groups will notice the target forms (Schmidt, 1990, 2001), consciously compare their own output

and the target forms (Schmidt & Frota, 1986), and modify their output (Swain, 1985). The comparison group participants who do not receive form-focused instruction will not improve syntactic accuracy because they are not explicitly directed to pay attention to the target forms.

Hypothesis 1b. All three groups (comparison, FFI, and FFI + PF) will improve syntactic complexity and oral fluency because task repetition has been found to lead to such improvements (Ahmadian, 2011; Bygate, 2001).

Hypothesis 2a. The FFI group will improve syntactic accuracy more than the comparison group because form-focused instruction explicitly directs them to pay attention to past tense forms. The FFI group participants will notice the target forms (Schmidt, 1990, 2001), consciously compare their own output and the target forms (Schmidt & Frota, 1986), and modify their output (Swain, 1985). Task repetition alone does not explicitly direct the comparison group participants to pay attention to the target forms and to notice errors in their use of the past tense forms. The FFI + PF group will improve syntactic accuracy more than the FFI group; more explicit form-focused instruction will lead to more syntactic accuracy development (DeKeyser, 1998; Goh & Burns, 2012; Norris & Ortega, 2000; Spada & Tomita, 2010).

Hypothesis 2b. There will not be significant differences in either syntactic complexity or oral fluency between the comparison group, FFI group, and FFI + PF group because the

form-focused instruction in this study does not target syntactic complexity or oral fluency development; all three constructs of syntactic complexity, syntactic accuracy, and oral fluency can improve from task repetition (Ahmadian, 2011; Bygate, 2001).

Hypothesis 3. The FFI + PF group will improve syntactic accuracy more quickly week by week than the FFI group and the comparison group because the FFI + PF group receives more form-focused instruction, which increases syntactic accuracy (Ahmadian & Tavakoli, 2011; Mochizuki & Ortega, 2008; Williams, 2005; Yuan & Ellis, 2003).

Hypothesis 4a. The FFI + PF group will increase classroom English speaking anxiety more than the FFI and comparison groups. The FFI group will increase more than the comparison group because it receives more form-focused instruction and teacher instruction could increase learner anxiety (Saito, 2004; Trang et al., 2013).

Hypothesis 4b. The FFI + PF group will increase speaking self-efficacy more than the FFI and comparison groups. The FFI group will increase more than the comparison group because form-focused instruction increases self-confidence (Van Batenburg et al., 2019).

Hypothesis 4c. The FFI + PF group will decrease in desire to speak English more than the FFI and comparison groups. The FFI group will decrease more than the comparison group because form-focused instruction can decrease learners' willingness to communicate (Van Batenburg et al., 2019).

Hypothesis 5a. The FFI + PF group will have higher self-assessments of syntactic accuracy than the FFI and comparison groups. The FFI group will have higher self-assessments of syntactic accuracy than the comparison group because more form-focused instruction can lead to greater syntactic accuracy gains (DeKeyser, 1998; Goh & Burns, 2012; Norris & Ortega, 2000; Spada & Tomita, 2010) and therefore higher self-assessments (Ma & Winke, 2019; Malabonga et al., 2005; Poehner, 2012).

Hypothesis 5b. The FFI + PF group and FFI group will not have higher self-assessment for syntactic complexity or oral fluency than the comparison group because form-focused instruction does not target improvements in syntactic complexity or oral fluency.

## Summary

In this chapter I reviewed the literature related to speech production and the conceptualization and measurement of second language speaking proficiency. In reviewing the measurement of second language speaking proficiency, I examined measures commonly used in previous studies, especially measures of syntactic complexity, syntactic accuracy, and oral fluency before examining the dynamic relationship among the CAF triad and the effects of task complexity on speaking. I also reviewed previous studies related to analytical ratings and learner self-assessment in second language learning. In the second part of this chapter, I reviewed the literature of communicative language teaching, task-based language teaching, focus on form and

focus on forms, and form-focused instruction before examining the effects of different form-focused tasks on speaking development. Finally, I reviewed the literature related to learner affective variables toward speaking including classroom English speaking anxiety, speaking self-efficacy, and desire to speak English. I concluded the chapter by outlining the gaps in the current speaking development literature and stated the five hypotheses that will be assessed in the present study.

In the next chapter I outline the methodology of the present study, beginning with a description of the research setting, participants, the interventions, the experimental groups, and the instruments used. I conclude the chapter with a detailed description of the procedures used, including descriptions of how the data were gathered and analyzed vis-vis the research hypotheses.

# CHAPTER 3

# METHODS

In this chapter I first present an overview of the research setting, the participants, the selection of target forms, and a description of the experimental groups. I then describe the instrumentation, research procedures, and analytical procedures used to answer the research hypotheses.

## Research Setting

This study was conducted in an EFL context in which the students had few or no opportunities to speak English outside of class. It was conducted at a prestigious national university in western Japan known for outstanding academic and research achievements. Students at this university have an average *hensachi* of 70. In Japan, *hensachi* is an indicator of students' academic ability and university entrance examination difficulty. *Hensachi* scores are usually published by large cram schools based on data collected from nationally administered mock university entrance examinations. Higher scores represent higher academic abilities and thus higher rankings for the universities. A *hensachi* score of 50 is the mean of the university entrance examination difficulty continuum. *Hensachi* scores have a standard deviation of 10. Students entering an academic institution with a *hensachi* of 70 are in the top 2.30% of the student population; in other words, they have higher academic standing than 97.70% of the Japanese university student population.

The students at the target university had a mean TOEFL-PBT score of 498.03 in 2017. Most students at this university have intermediate- or advanced-proficiency English reading and listening skills although a large disparity exists between the top and bottom achieving students. All of the participants were required to enroll in two English classes per week in both the spring and fall semesters of the first two academic years; thus, in the first four semesters, students enroll in a minimum of eight English courses. Although the students could choose their English courses, their choices were restricted by department timetable availability; therefore, students with the same majors were usually enrolled in the same English courses. The courses are held for one academic semester, so the students had different instructors in the spring and fall semesters. Each course was made up of fifteen 90-minute classes that met once a week. English classes were taught by full-time and part-time Japanese and non-Japanese instructors; Japanese instructors represented a larger proportion of the faculty members. The mandatory English courses did not follow a standardized curriculum, so different instructors at the target university used different syllabi, educational objectives, and pedagogical materials. This study was conducted from the fourth week (May, 2015) of the semester in order to let the participants become familiar with their university courses and classrooms. The study concluded at the end (July, 2015) of spring semester.

## Participants

The participants were 104 first-year students (80 male and 24 female students) whose ages ranged from 18 to 20. They were majoring in law, economics, engineering,

letters, or human sciences. All participants were native speakers of Japanese who had received six years of mandatory English education in Japanese junior and senior high schools. They came from five intact communication-based English classes I taught. The participants were informed of the purpose of the present study and were asked to sign a consent form in Japanese (see Appendix A for the original Japanese form and Appendix B for an English translation). The Japanese consent form was distributed to the participants before the beginning of the study. The participants were informed that their participation was voluntary and not related to their class grade.

All five classes were required English courses. The students could choose mandatory English courses focused on reading, writing, listening, or speaking. The same syllabus was standardized for these five classes to eliminate the potential effects of students engaging in different tasks. The syllabus was focused on the theme of global studies, so various contemporary topics such as international terrorism and poverty were introduced each week. The students engaged in task-based activities, including watching short videos, pair work, small group discussions, and group presentation tasks. Teacher-centered, form-focused instruction was minimized; the teacher's role was primarily that of a facilitator.

Although all the participants had received the same number of years of English education in junior and senior high school, they had different experiences with English in other settings (Table 2). Table 2 summarizes information about the participants gathered with the background questionnaire (see Appendix C), which was used to gather information about the participants' English language proficiency, previous English

learning experience, and exposure to English outside of class. All participants were asked to use *Surveymonkey* to complete the background questionnaire. In the background questionnaire, the participants were also asked to indicate their *hensachi* and scores of any previously completed English proficiency tests. Most of the participants had engaged in some English-speaking activities in junior and senior high school. Approximately 14.00% attended private English conversation schools and about 9.00% had engaged in English-related extracurricular activities such as going to an English café provided by the university or having international student friends. Finally, approximately 22.00% of the participants had been outside of Japan for holidays or short study aboard programs.

Table 2. *Participants' Past Experience with English Activities*

| | Yes (%) | No (%) |
|---|---|---|
| Attended private English conversation school | 14 | 86 |
| Visited overseas (including studying aboard) | 22 | 78 |
| English-related extra-curricular activity | 9 | 91 |

## Selection of Target Forms

In the present study, three English past tense forms—simple past tense, past continuous tense, and past perfect tense—were selected as the target grammatical forms. Past tense forms were chosen because they are necessary for retelling stories that happened in the past.

Deciding on appropriate target forms for the intervention was important because different syntactic forms are differentially sensitive to the effects of form-focused instruction (Nassaji, 2000). Ellis (1995) suggested that the selection of the target forms should be based on problematicity and learnability. Problematicity, which refers to the

difficulty that target learners have with particular structures, can be determined by examining samples of the target learners' production (Ellis, 1995) or be determined intuitively based on the teacher's perception of the target learners' language needs (Nassaji, 2000). Learnability refers to the level at which learners can integrate target syntactic forms into their linguistic system notwithstanding the difficulty of determining which forms learners are ready to learn (Ellis, 1995). In the present study, simple past, past continuous tense, and past perfect tenses were chosen as the target of intervention for three reasons.

First, Japanese learners experience difficulty using simple past tense accurately even at intermediate- or advanced-proficiency levels despite the fact that past tense forms are often introduced in the early stages of foreign language instruction (Ellis et al., 2006). Past tense forms are considered to be moderately difficult for EFL learners (Xu & Lyster, 2014). In addition, simple past, past continuous, and past perfect tenses can be difficult for second language learners to produce when speaking, even when they have explicit knowledge of the form (Doughty & Varela, 1998; Ellis et al., 2006). As Ellis et al. stated, "Learners, especially Asian learners tend to produce errors of past tense because of their phonological difficulties in producing consonant clusters with final [t] or [d]" (p. 351).

Second, almost all of the participants had explicit knowledge of the three past tense forms as all Japanese university students (except with rare exceptions) study all three forms in secondary school. I confirmed the students' explicit knowledge of these three past tense forms from their written homework prior to starting the study. In addition, the participants in the pilot study made frequent past tense errors even though

they stated that they had studied the three tense forms in high school. Many participants in the pilot study knew the difference between present and past tense forms but they were unable to produce them accurately when speaking.

Third, past tense forms can be further categorized into regular and irregular forms; this distinction allows for the comparison of the effect of form-focused instruction on both rule-based (regular past tense) and item-based (irregular past tense) forms (Yang & Lyster, 2010). Regular and irregular past verbs have different forms but similar functional and semantic values. Therefore, changes in the use of regular and irregular forms are not the result of the different communicative value of the target forms. Learners experience difficulties using past tense and past progressive and past perfect aspects accurately because they have difficulty judging when the *-ed* form or an irregular form is used (e.g., *eated* vs. *ate*) or the interchange the simple past and past participle forms (e.g., *ate* vs. *eaten*).

## Intervention Groups and the Comparison Group

In this quasi-experiment study, five classes were randomly assigned to a comparison condition group or to one of the two treatment groups, the form-focused instruction group (FFI) or the form-focused instruction and peer feedback group (FFI + PF). Teaching logs were recorded in each class during the experiment period to ensure that similar activities were provided to all three groups, with the primary difference being the opportunities for explicit form-focused grammar instruction. Two intervention groups were included in order to investigate the difference between the effects of giving more

and less explicit form-focused instruction on oral proficiency development. Both the comparison group and the two intervention groups engaged in the same narration task at the beginning of every class. The narration task lasted approximately three minutes, which included one minute of planning time and two minutes of speaking time. While some researchers have used longer planning time (e.g., Skehan & Foster, 1997; Yuan & Ellis, 2003), I decided to use one minute of planning time, which is similar to other speaking development studies (e.g., Mehnert, 1998; Wigglesworth, 1997; Wigglesworth & Elder, 2010). One minute of planning time controlled for the effects of pre-task planning. Planning can lead learners to pay attention to linguistic form (Mochizuki & Ortega; Sangarun, 2005) because it leads them to improved working memory functioning and a better conceptualization of the task before the performance (Ellis, 2009a). In each weekly session, the participants in all three groups repeated the speaking narration task twice. I did not help or encourage the participants to self-correct their errors. The two intervention groups received a form-focused intervention between the first and second narration while the comparison group did not receive any intervention. Participants in the intervention groups were not taught the use of past tense forms because all Japanese learners study these forms in high school. Instead, their awareness and ability to monitor their use of past tense forms was raised via the interventions. As indicated by previous studies, self-monitoring is a crucial component for second language speaking development (Bygate, 2009; Kormos, 2005; Levelt, 1989). Each week during the seven-week intervention, intervention group participants received instruction on two past tense forms so that they could compare the two forms: (a) simple past and past continuous, (b)

past continuous and past perfect, and (c) simple past and past perfect (See Table 3). The intervention in Week 1 included the simple past tense and past continuous because the natural order of English past tense acquisition goes from simple past to past continuous, to the past perfect tense (Bardovi-Harlig, 2000). The intervention in Week 7 included all three target forms.

Table 3. *Grammar Focus for Weekly Intervention*

| Week | Focus |
|---|---|
| 1 | Simple past and past continuous |
| 2 | Past continuous and past perfect |
| 3 | Past perfect and simple past |
| 4 | Simple past and past continuous |
| 5 | Past continuous and past perfect |
| 6 | Past perfect and simple past |
| 7 | Simple past, past continuous, and past perfect |

**Form-Focused Instruction Group**

The form-focused instruction group received approximately 10 minutes of explicit teacher-led grammar instruction between the first and second narration for seven weeks on three past tense forms: simple past, past continuous, and past perfect. I provided 10 minutes of grammar instruction to the participants based on the results of a pilot study conducted on the effects of form-focused instruction on speaking development. In the pilot study, participants in the two intervention groups were given either a form-focused instruction or self-review time each week. The weekly intervention took approximately 90-seconds. The pilot study results indicated that the differences between the intervention groups and the comparison group produced no significant results in terms of mean length of T-unit, mean length of clause, clause/T-unit ratio, percentage of error-free T-units,

speech rate, and self-repair; therefore, I concluded that a 90-second intervention was insufficient to trigger speaking performance improvements. Furthermore, the effects of the form-focused instruction on speaking proficiency development might have been suppressed due to the short intervention; therefore, the weekly intervention time in the present study was increased to 10 minutes. Each week, the grammar intervention included the following three components.

First, the FFI group participants received whole-class grammar instruction, which was a short conscious-rising activity for approximately two minutes. The target structures were written on the blackboard with three example sentences (see Appendix D). The participants' attention was directed to the target forms by having them repeat the example sentences aloud while I pointed at the target features with a pointer. Differences between the target tenses and their uses were also deductively explained on the board. Finally, the participants were reminded to avoid making careless mistakes with target past tense forms during the speaking tasks. During the intervention, I raised the participants' awareness of the past tense forms they had learned in secondary school. Therefore, the participants' task was to monitor their use of the past tense, not acquire new forms.

Second, the participants completed an individual grammar worksheet on the target past tense forms, which was written output practice that took approximately five minutes. The grammar worksheet included questions in which the participants filled missing verbs in blanks to complete sentences, filled missing verbs in blanks to complete sentences by changing present tense verbs into affirmative and negative past tense forms, and filled sentences or partial sentences into blanks to complete paragraphs (see Appendix E for an

example grammar worksheet). The grammar worksheets were included as a form-focused instruction component because they directed the learners' attention to the target linguistic forms (VanPatten, 1993). Grammar worksheets provided the participants the opportunity to internalize the target forms via repetition and practice, so they were more likely to access the target forms during oral production. The grammar worksheets were adapted from a commercial grammar textbook, *English grammar in use: A self-study reference and practice book for intermediate learners of English* (4th ed.) (Murphy, 2012). This textbook was chosen because it was considered a suitable grammar textbook for the participants' proficiency level after consulting with other teachers and piloting the textbook with students attending the target university. In the pilot study, the participants completed questions from the target grammar textbook while explaining verbally how they reached their answers. In addition, the participants in the pilot study were asked to comment on the difficulty of the questions in the grammar textbook. The instructions and explanations on the worksheets were not translated into Japanese because the participants in the pilot study commented that the original instructions and explanations in English were easily comprehensible. After the participants completed the grammar worksheets individually, I wrote and explained the answers for the questions on the grammar worksheets on the blackboard while the students listened and checked answers on their own grammar worksheets.

Third, the FFI group participants completed a pair oral output practice tasks that were focused on the target forms for approximately two minutes. The participants were paired with a partner. A list of ten verbs related to the content of each week's narration

was shown on a computer screen in their base form (see Appendix F for an example). The participants chose one verb from the computer screen and verbally made three sentences using the same verb conjugated as simple past, past continuous, and past perfect tenses (e.g., *I slept yesterday*, *I was sleeping yesterday*, and *I had slept*). In pairs, the participants verbally produced three sentences while their partner listened and then the pair of participants switched roles. This process was repeated with other verbs until the end of intervention time. The form-focused instruction group participants spent approximately 10 minutes discussing global topics in small groups instead of receiving the peer feedback intervention, which was given only to the form-focused instruction and peer feedback group.

**Form-Focused Instruction and Peer Feedback Group**

The form-focused instruction and peer feedback (FFI + PF) group received approximately 20 minutes of intervention prior to the cartoon narration each week; the intervention time given to this group was twice the amount of intervention time given to the FFI group. The intervention for the FFI + PF group included the same 10 minutes of grammar instruction provided to the FFI group followed by an additional 10-minute peer feedback activity. The grammar instruction and peer feedback components occurred consecutively without other tasks in between.

For the peer feedback activity, the FFI + PF group participants listened to another participant's narration and provided feedback for improving the peer's syntactic complexity, syntactic accuracy, and oral fluency. One week before the first peer feedback

activity, the participants received training on how to conduct peer feedback. In the training session, I explained the purpose of peer feedback and demonstrated how to give specific constructive comments for improving the three speaking constructs. In addition, all participants practiced giving feedback by listening to one model narration recording. They wrote comments on a peer feedback comment sheet (see Appendices G and H for Japanese and English versions of the comment sheet; see Appendix I for an example of completed comment sheet). The participants were asked to share and discuss the comments they made with other participants in small groups.

The weekly peer feedback activity included the following steps. After the participants recorded their first narration individually on the computer, they worked with a partner to complete peer feedback. Each participant listened to their partner's recorded first narration performance and provided comments in English, Japanese, or a combination of the two languages. The FFI + PF group participants were encouraged to give specific examples as part of their comments. After receiving feedback from their partner, the speakers read the feedback given by the reviewers while listening to their first narration again. The peer feedback comment sheets were collected before the participants began their second narration so that the participants would not read directly from the prompts given on the feedback sheets. After the peer feedback comment sheets were collected, the FFI + PF group participants narrated the cartoon again.

### Comparison Group

The comparison group carried out the same weekly narrative task as the intervention groups except the comparison group participants did not receive any instruction before the weekly narration task. Instead of receiving form-focused instruction, the comparison group participants spent approximately 20 minutes discussing global topics in small groups. Table 4 summarizes the amount of time allocated for planning and narration for the three groups. It also summarizes the treatment given to the two intervention groups.

Table 4. *Weekly Interventions for FFI and FFI + PF Groups*

| Group | Intervention | Time (min.) |
|---|---|---|
| FFI | • Whole class blackboard grammar explanation | 2 |
| (*n* = 41) | • Individual grammar worksheet | 3 |
| | • Pair speaking activity using target forms | 5 |
| FFI + PF | • Whole class blackboard grammar explanation | 2 |
| (*n* = 38) | • Individual grammar worksheet | 3 |
| | • Pair speaking activity using target forms | 5 |
| | • Peer corrective feedback | 10 |
| Comparison | • No intervention was given | |
| (*n* = 25) | • Participants did other discussion tasks | |

*Note*. FFI (Form-focused instruction) group participants were engineering majors, FFI + PF (Form-focused instruction and peer feedback) group participants were humanities majors, and comparison group participants were law and letter majors.

## Instrumentation

A mixed method design focused predominantly on quantitative measures with the support of qualitative data was employed in this study. Data were collected from five sources: a background questionnaire, a four-picture cartoon narration, the Self-Assessment of Speaking Task Questionnaire, human analytical ratings, and the English

Speaking Learner Affect Questionnaire. All instruments were piloted prior to the start of the study with students attending the same university as the target participants. In addition, the data collection process was piloted during one academic semester with a similar number of students at the target university. The questionnaires were reviewed by three English teachers (including me) to eliminate ambiguity before they were translated into Japanese by a Japanese research assistant. The translated Japanese items were back translated into English by another Japanese research assistant and then compared with the original English items to avoid discrepancies in meaning. The questionnaires were then piloted with three students from the target university to confirm the time required for questionnaire distribution, completion, and collection.

**Four-Picture Cartoon Narration**

Cartoon narration was chosen as the speaking task of the present study because the semi-structured narrative standardizes content and task difficulty (Kormos & Dénes, 2004; Mora, 2006). This instrument was adapted from the *Eiken* speaking test, which is one of the most widely used English-language testing programs in Japan for assessing English proficiency. Cartoon narrative task equivalent to *Eiken* Grade Pre-1 was chosen because the pilot study indicated that this level was suitable for the target participants. In addition, the *Eiken* website (https://www.eiken.or.jp/eiken/en/grades/) stated that Grade Pre-1 is suitable for examinees who have sufficient proficiency to understand and use English, and to participate effectively in social, professional, and educational situations. The task was to narrate the story of the four-panel cartoon that had four stages in

chronological order: introduction, goal, conflict, and resolution (conclusion). Picture 1 sets up the context of the story with the main characters, Picture 2 develops the story where characters attempt to achieve some goals, Picture 3 presents the climax of the story where the characters encounter conflicts, and finally the conflicts are resolved in Picture 4. The four-picture cartoon is considered to be suitable for the target participants because Japanese learners are familiar with *kishōtenketsu*, which is the classic Japanese narrative made up of four parts: introduction, development, conflict, and conclusion.

The participants were given one-minute of planning time before they were asked to narrate the cartoon for two minutes. Eight cartoons were used in this study (See Appendix J for an example). The same cartoon was used for the pretest and posttest, and seven different cartoons were used in the seven intervention weeks (one cartoon per week). The cartoons were taken from a commercial *Eiken* preparation textbook, *Eiken junikyu nijishiken mensetsu kanzen yosou mondai* (Akao, 2011). The cartoons from the commercial textbook were piloted for content difficulty. Eight cartoons were selected from a set of 14 cartoons based on comments made by 20 students interviewed in a pilot study. The pilot study deleted six out of the 14 cartoons that were considered to have either ambiguous and (or) difficult story structure by the student interviewees.

**Self-Assessment of Speaking Task Questionnaire**

The Self-Assessment of Speaking Task Questionnaire was developed to measure the participants' assessment of their oral narrative performance on the pretest and posttest in terms of syntactic complexity, syntactic accuracy, and oral fluency (See Appendix K

for original Japanese version and Appendix L for the English translation). Previous studies on speaking self-efficacy (e.g., Apple, 2011; Leeming, 2017) have mostly used *can-do* statements to measure learners' perception of their overall speaking proficiency. There is no other study to my knowledge that investigated learner's assessment of their own speaking task performance using CAF measurements. In this study, the participants were asked to assess their own speaking performance of the cartoon narrative task immediately after narrating the pretest and posttest cartoon. Participants were instructed to assess their specific performance of the narrative task rather than assessing their general speaking proficiency.

Questionnaires are considered a suitable approach to measuring self-assessed proficiency because they offer a reflective process for learners to identify gaps in their knowledge (Alderson, 2005; Little, 2005; Trofimovich et al., 2016). The questionnaire is in a format that most participants are familiar with, so they are likely to assess their performance more precisely and appropriately. Questionnaires are also considered effective for administration, data collection, and analysis (Dörnyei, 2003); however, there are also limitations with using questionnaires; some researchers have claimed that questionnaires can produce unreliable and invalid data when ineffectively constructed (Dörnyei, 2003). Respondents who have more interest in the subject matter might answer some items more precisely while ignoring other items. Some respondents might misunderstand some items due to poor design or ambiguous wording. Finally, there is the possibility of fatigue when the same questionnaire is administrated more than once.

The Self-Assessment of Speaking Task Questionnaire items were developed by referring to definitions and measurements of speaking proficiency in terms of syntactic complexity, syntactic accuracy, and oral fluency from the studies reviewed in Chapter 2 (e.g., Foster et al., 2000; Iwashita, 2010; Skehan & Foster, 1999). Because no other study to my knowledge has used similar questionnaire items to examine learners' assessment of their speaking task performances using the CAF triad, a number of pilot studies were conducted to examine item functioning with first-year students from the target university. A total of 19 items were developed to measure three latent constructs: five items for syntactic complexity, seven items for syntactic accuracy, and seven items for oral fluency. A six-point Likert scale (1 = *Strongly Disagree*, 2 = *Disagree*, 3 = *Slightly disagree*, 4 = *Slightly agree*, 5 = *Agree*, 6 = *Strongly agree*) was used as it was able to extract more information out of questionnaire items than dichotomous *yes/no* items (Linacre, 2002b). The six-point rating scale was selected to increase measurement precision (Nemoto & Beglar, 2014).

**Human Analytical Ratings**

The pretest and posttest recordings for all participants were assessed by human raters because human raters offer a subjective impressionistic judgment of overall oral performance that complements the objective CAF measures (Nitta & Nakatsuhara, 2014). Analytical rating was chosen over holistic rating because it offers an examination of specific subcomponents of oral proficiency (Xi & Mollaun, 2006). Analytical ratings also have higher reliability and construct validity than holistic ratings (Ling et al., 2014).

In the pilot study, I examined the suitability of the original *Eiken* speaking test scoring criteria on a 1-5 scale by rating 59 participants. According to the original *Eiken* scale, speakers scoring a 4 or 5 can *use a wide range of vocabulary appropriately and show accuracy in grammar and word usage*. However, no participant chose a 4 or 5 on the scale. Most of the participants in the pilot study selected either 1 or 2, with a few participants choosing 3, which indicates that *errors in grammar and word use occur but do not really impede communication*. In addition, the original *Eiken* scoring criteria divides speaking proficiency into three combined construct sections: coherency and content, pronunciation and intonation, and the third section of vocabulary, grammar, and word usage. After consulting with other English teachers from the target university and second language speaking development researchers, it was decided that the original *Eiken* criteria were unsuitable for the purposes of the present study because they are focused on constructs such as content, lexis, and organization that were not the focus of this study. Therefore, I adopted the rating criteria used by Nitta and Nakatsuhara (2014), who focused on assessing complexity, accuracy, and fluency (see Appendix M); their criteria were modified from Iwashita, Elder, and McNamara (2001). In the present study, content was not included on the analytical rating rubric because the semi-structured cartoon narrative task standardized the content (Kormos & Dénes, 2004; Mora, 2006). Because the content was decided in advance, it was not necessary to assess it (Sato, 2014). The original 0-8 scale (Nitta & Nakatsuhara, 2014) was converted to a 1-5 scale for the present study in order to decrease raters' cognitive load (Nemoto & Beglar, 2014).

Three raters including myself rated the participants. The other two raters were Japanese university English teachers with more than ten years of English teaching experience who were both doctoral students in a Ph.D. program in Japan. One rater is a native speaker of English and the other rater is Japanese who has native-level English proficiency. The Japanese rater has a TOEIC score of 970 and has been teaching English in Japanese universities for more than ten years. All raters have over 15 years of teaching experience and are familiar with conducting analytical ratings for teaching and research. The raters were given an orientation on how to use the rating scale through meetings. During the rater orientation, the raters were given rating guidelines (See Appendix N) containing information about the purpose of the assessment, the assessment criteria, and the rating procedures as well as background information about the participants, including their English proficiency and the speaking tasks they completed in the classroom during the research period. The raters were asked to review the grading criteria and comment on any ambiguous wording of the criteria. Pilot ratings were conducted among the three raters; the raters were provided with 15 recordings of cartoon narrations from a different group of participants in the target university. The raters, who were asked to rate the participants' speaking proficiency with no regard to producing a normal distribution, were asked to listen to each recording only once and then provide three analytical scores for complexity, accuracy, and fluency using the rating criteria. The scores were compared and minor differences in the scoring were identified and discussed. In order to prevent fatigue, the raters were asked to take a short break between narrations.

The pretest and posttest speaking performances of all 104 participants (104 participants x 2 performances = 208 total performances) were independently rated by the three raters. The recordings were coded from 1 to 208 and were given to the raters using *Googledoc*. The raters listened to each monologue once, rating it on the three analytical criteria of complexity, accuracy, and fluency. They were asked to submit their scores within one month; thus, they could assess the recordings at their own pace. Each rater rated approximately 50 participants with at least 20 participants overlap-rated by two raters. Overlap-rating was used to ensure enough linkage between all of the elements and parameters in the model (Linacre, 2011). In the pretest, Rater 1 rated 48 participants (Participant 1 to Participant 48), Rater 2 rated 37 participants (Participant 67 to Participant 104), and Rater 3 rated 49 participants (Participant 30 to Participant 79). In the posttest, Rater 1 rated 51 participants (Participant 53 to Participant 104), Rater 2 rated 30 participants (Participant 40 to Participant 70), and Rater 3 rated 50 participants (Participant 1 to Participant 51).

**English Speaking Learner Affect Questionnaire**

The English Speaking Learner Affect Questionnaire was developed to measure the participants' affect toward speaking English in EFL classrooms on the pretest and posttest in terms of classroom English speaking anxiety, English speaking self-efficacy, and desire to speak English (see Appendix O for original Japanese version and Appendix P for the English translated version). Learner affect can positively and negatively influence language performance and proficiency (Yashima, 2002); therefore,

understanding learner affect toward speaking English can contribute to understanding their speaking performance and development. Up to now, numerous researchers have used questionnaires to measure affective variables such as anxiety, self-confidence, and willingness to communicate (e.g. Dörnyei, 2005; Horwitz et al., 1986; MacIntyre and associates, 1991, 1997, 1998, 1999, 2001, 2003; Yashima, 2002).

This questionnaire consists of 40 items measuring three constructs: classroom English speaking anxiety (18 items), English speaking self-efficacy (11 items), and desire to speak English (11 items). The items were adapted from Apple (2011), Horwitz, Horwitz, and Cope (1986), and Yashima (2002). Some items were rephrased to describe the learning context of the target university. The following six-point Likert scale was used: 1 = *Strongly disagree*, 2 = *Disagree*, 3 = *Slightly disagree*, 4 = *Slightly agree*, 5 = *Agree*, 6 = *Strongly agree*. The six-point rating scale was selected for this questionnaire to ensure a degree of measurement precision (Nemoto & Beglar, 2014).

## Procedures

Procedures for the nine-week research period with the seven-week intervention and data collection methods are described below. The present study was conducted in my English classes from May to the last week of July. Before the start of the research, four weeks of class time were allocated at the beginning of the semester to familiarize the participants with the research setting, my instructional style, and with other participants in the class. The research period included one pretest week, seven weeks of interventions, and one posttest week (see Figure 3). One practice narration was conducted one week

before the start of the research period where all participants practiced narrating an equivalent four-picture cartoon in Japanese. Narrating the cartoon in Japanese provided all participants with the opportunity to become familiar with the four-picture cartoon; they practiced planning, managing, and allocating enough time for each picture in the cartoon. The participants also had the opportunity to become familiar with using *Movietelecom*, the recording software installed on their individual computers. The cartoon used for Japanese narration practice was not used in the rest of the study.

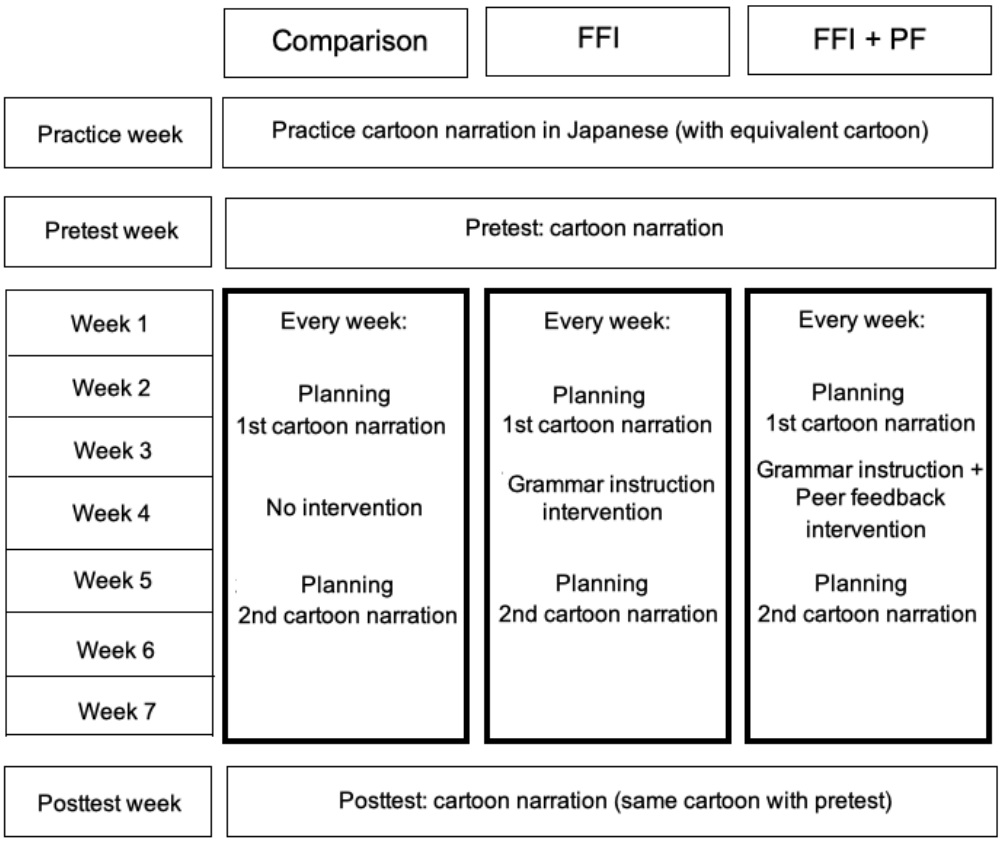



*Figure 3*. The design of present research.

All participants narrated and recorded the four-picture cartoon pretest in English. The participants were told to speak as much as possible during the narration time of two minutes. They narrated the pretest cartoon twice consecutively with one-minute planning time before both the first and second narrations. After completing the pretest, the participants completed the Self-Assessment of Speaking Task Questionnaire and the English Speaking Learner Affect Questionnaire using the online *SurveyMonkey* software.

Seven consecutive weeks of intervention were provided one week after the pretest. The participants narrated a different four-picture cartoon every week during seven weeks of interventions. Each week, the participants in the FFI group and FFI + PF group received their interventions between the first and second narration. The participants in the comparison group did not receive any intervention between the first and second narration each week. All participants narrated the same four-picture cartoon used for the pretest on the posttest. After completing the posttest, all participants completed the Self-Assessment of Speaking Task Questionnaire and the English Speaking Learner Affect Questionnaire again using online software *SurveyMonkey*.

## Data Deletion

The data gathered from some participants were deleted before conducting the statistical analyses. First, participants who did not complete both the pretest and posttest as well as the participants who missed more than one of the seven intervention narration sessions were excluded from the analyses because the present study was designed to examine the longitudinal effects of form-focused instruction on speaking. The data of 14

participants were deleted. Second, data collected from international students ($n$ = 11) were also deleted from the analyses because these students might have had more opportunities to speak English outside of the classroom than the Japanese participants and they also had higher English-speaking proficiency. The data of the remaining 104 participants were used for the analyses in the present study.

## Data Analyzing Procedures

The oral data were manually transcribed and pruned by myself because I was familiar with the English proficiency level and use of English grammar of the target participants from teaching at the university. Before the analysis, I practiced coding the audio transcripts in order to minimalize inconsistencies because past researchers (e.g., Iwashita et al., 2008) have commented that it is difficult to maintain consistency in coding. The transcribed and pruned data were manually analyzed by me to measure the participants' speaking proficiency in terms of syntactic complexity, syntactic accuracy, and oral fluency. Oral fluency was analyzed first because self-repairs were deleted before conducting the complexity and accuracy analyses. Pausing measures for the oral fluency component were not conducted because of the large *N*-size of 104 participants. Inter-rater reliability was calculated with approximately 10.00% of the oral data that had been checked by a research assistant who was a qualified Japanese English teacher and a post-graduate student in applied linguistics. The research assistant had worked together with me on previous CAF studies and was familiar with CAF measurements and coding.

Disagreements in coding occurred with less than 0.50% of the data; the disagreements were discussed and solutions were negotiated.

Statistical Package for Social Sciences (SPSS) version 25 was used to analyze the six CAF measurements. The gain scores (posttest - pretest) of the participants from the three groups were analyzed. One-way repeated-measures univariate ANOVAs were conducted to examine within-group differences of the three groups on the six CAF measures. A one-way MANOVA followed by six one-way ANOVAs were used to examine between-group differences.

**Transcription and Pruning**

I orthographically transcribed the recordings of the pretest and posttest narrations using guidelines adapted from Ortega, Iwashita, Rabie, and Norris' transcription guidelines (1999). The total speech time (in seconds) of each narration was calculated by deleting the initial silence before the first sound and the silence after the last sound as well as utterances not related to the narration (e.g., *Thank you for listening*, *I start speaking*, or *That's all*). In addition, utterances at the end of the narration that were only partial sentences with insufficient content to create meaning were also deleted (e.g., *She felt it is so aaaaaa*).

The transcriptions were pruned by referring to previous pruning literature (Kawauchi, 2005; Thai & Boers, 2016). Silent pauses under three seconds were included in the analysis because they are part of natural speech, whereas silent pauses over three seconds were deleted from the analysis (Riggenbach, 1991). Filler pauses in both English

(e.g., *ehm*, *uh*, *err*, *ee*, *ah*, *um*) and Japanese (e.g., *ano*, *eto*, *eeh*) were deleted because they are not part of the content of the speech (Freed, 2000; Iwashita, 2010). Filler pauses between two identical words (e.g., *watch seal ano seal on TV*) were deleted.

**Selection of CAF Measurements**

I analyzed the pruned transcription in terms of syntactic complexity, syntactic accuracy, and oral fluency using six CAF measures (Table 5) adapted from previous studies (e.g., Ellis 2009a; Ellis & Barkhuizen, 2005; Iwashita et al., 2008; Norris & Ortega, 2009). Two measurements were employed to assess syntactic complexity: mean length of T-unit, clause/T-unit ratio. Two measurements were employed to assess syntactic accuracy: percentage of error-free T-units and percentage of accurate past tense verb usage. Finally, two measurements were employed to assess oral fluency: speech rate and self-repair. The fluency analyses were conducted before the complexity and accuracy analyses. Self-repairs were coded and pruned before speech rate was calculated because the participants who accurately repaired utterances were considered to have produced accurate target features (Iwashita et al., 2008). For instance, "one day a girl was watching watching watching TV program on sea animals seals" was pruned to "one day a girl was watching TV program on seals." Lexical measurements were not included in the present study even though some researchers have argued that lexis needs to be added to the CAF model (Lennon, 2000; Skehan, 2009b, 2018; Skehan & Foster, 2007; Vercellotti, 2017). While lexical access and retrieval plays an important role in speech production, differences in lexical range are not as apparent in non-native speakers as in native

speakers (Skehan, 2009b). In addition, lexical analyses were excluded from the present study because vocabulary in narrative telling strongly depends on the structure and storyline of the prompt (de Jong & Vercellotti, 2016); therefore, lexical diversity in the present study was limited because the storyline of the four-picture cartoon narrative task was predetermined.

Table 5. *CAF Measurements*

| Features |
|---|
| Syntactic complexity |
| 1 Mean length of T-unit |
| 2 Clauses per T-unit ratio (total number of dependent clauses divide total number of T-units) complexity by subordination |
| Syntactic accuracy |
| 1 Percentage of error-free T-units |
| 2 Percentage of accurate past tense verb usage |
| Oral fluency |
| 1 Speech rate |
| 2 Self-repair |

T-unit analysis was chosen as the main unit of analysis even though more speaking researchers have shifted to AS-unit based analyses (Foster et al., 2000). Foster et al. argued that the AS-unit is more appropriate for analyzing spoken data because it is a "single speaker's utterance consisting of an independent clause or sub-clausal unit, together with any subordinate clause(s) associated with it (p. 365). However, T-unit analysis was chosen for the present study because it can be used interchangeably with AS-units to include incomplete and partial clauses (Foster et al., 2000). Even though T-units and AS-units are defined similarly (Foster et al., 2000), AS-units were not used in the present study because AS-units take into consideration intonational patterns and

pauses (Foster et al., 2000); these features were excluded from the present study. T-unit analysis, on the other hand, is considered sufficient for monologues especially for narrative analysis, as narratives do not include turn-taking elements required in conversational analysis.

**Syntactic Complexity Analysis**

In the present study, the complexity analysis was focused on syntax; other complexity elements such as content, vocabulary, and pragmatics were excluded. A T-unit was defined as one independent clause plus any dependent clauses that were contained in or attached to it plus any partial or incomplete sub-clauses. An independent clause was defined as a clause that can stand alone as a complete sentence. A dependent (subordinate) clause is defined as a clause that contains a subject and a verb that cannot stand on its own. Partial sentences, such as, *I eat,* or verbless utterances such as, *yes please,* were considered to be partial sub-clauses and were not included in the complexity analysis count. One T-unit contained multiple dependent clauses and partial sub-clauses.

First, a general measure of syntactic complexity was gathered by calculating the mean length of T-units, that is, the average number of words in each T-unit. Second, subordination complexity was measured by calculating the clauses/T-unit ratio. Some past researchers used the total number of different clauses (i.e. dependent clauses + independent clauses, or dependent clauses + independent clauses + sub-clausal units) (e.g., Ahmadian, 2012), whereas others used only the number of dependent clauses (e.g., Michel, Kuiken & Vedder, 2007) as the numerator. An examination of previous studies

using subordinate complexity measurements revealed that they did not indicate clearly which type of clause(s) was (were) included in the subordinate complexity calculation. In the present study I counted both independent and dependent clauses and used their sum as the numerator for calculating subordinate complexity. Dependent clauses were placed in brackets [ …] (Foster et al., 2000). For example, *they thought that* [*they should move to a larger place* [*because they have trouble living there.*]] included two dependent clauses. Three types of dependent clauses were included in the count: adverb clauses, noun clauses, and adjective clauses.

**Syntactic Accuracy Analysis**

The present study was focused on two types of accuracy analyses, global errors and specific errors associated with past tense usage. Global accuracy was examined by calculating the percentage of error-free T-units, which concerns the number of accurate T-units against the total number of T-units produced. Although some researchers (e.g., Polio, 1997) have raised concerns regarding the calculation of percentage of error-free T-units, as it does not consider the severity of errors or the number of errors in each identified error T-unit, it has nevertheless been one of the most commonly applied accuracy measurements in previous studies. Percentage of error-free T-units analysis in this study included the following grammatical errors: verb tense errors (e.g., Tomorrow, she joined eco tour and enjoy watching seals), verb aspect errors (e.g., She watching TV program yesterday), subject-verb agreement errors (e.g., She watch_ TV program), plural markers (e.g., There are two apple in the basket), prepositions errors, word order errors

(e.g., One day TV program a girl was watching), the omission of pronouns (e.g., Two months later, _watch program on TV), and *be* verb errors (e.g. many animals _ in danger) (Iwashita et al., 2008). Finally, T-units whose meaning was not comprehensible, were also counted as errors. Unlike previous studies in which both the percentage of error-free T-units and percentage of error-free clauses were calculated, only the percentage of error-free T-units was included in the present study because calculating both the percentage of error-free T-units and the percentage of error-free clauses would have generated overlapping, redundant results.

Several exclusions were made after consulting with other English teachers at the target university over the proficiency level of the participants and from examining the results of pilot studies. First, article errors were excluded because they are considered one of the most difficult linguistic features for Japanese learners of English to acquire. Past studies have indicated that Japanese learners have difficulty with the use of English articles because Japanese does not have articles; therefore, many Japanese learners tend to omit articles or make errors using them (Muranoi, 2000). Students need to be led to specifically focus on this form in order to improve their article performance. Because the present study was not focused on articles, it was expected that the participants would not improve their use of articles; thus, article errors were excluded from the error-free T-unit analysis. Second, morphological errors were excluded from the global error analysis because they are considered less serious and thus local (e.g., *Tours* are in danger instead of *tourists*), whereas syntactic errors are considered more disruptive and thus global (Tonkyn, 2012, p. 225). Third, lexical errors were also excluded (e.g., *She went to travel*

*shop yesterday*) unless it was a usage error between a verb, adjective, or noun (e.g., *she is interesting in human pollution issues*). The final exclusion from the error-free T-unit analysis was the word *tourist* and its plural marker *tourists*. This error occurred frequently and was considered a lexical error rather than a plural marker error.

For the specific error analysis, three aspects of past tense were investigated: simple past, past continuous, and past perfect. Past perfect continuous was excluded from the specific error analysis because it was decided after consulting with other English teachers that past perfect continuous form is not commonly used in spoken English discourse. The percentage of accurate past tense verb usage was calculated by counting the total number of past tense uses and the respective number of accurate forms. This analysis included present and past tense errors (e.g., *Yesterday a girl watch program on TV*), regular and irregular past tense errors (e.g., *Yesterday a girl eated lunch by herself*), as well as the production of inappropriate words that changed the nuance of the sentence were counted (e.g., Yesterday a girl washed program on TV). In addition, passive forms were included into counting the accurate use of past tense verbs, whereas *be* verbs acting as the main verb in the sentence (e.g., *She is a student*) were excluded because they were not the target of form-focused instruction intervention.

First, I counted the frequency of simple past tense forms used, which included the number of appropriate uses of the simple past tense. For example, places where the participants should have used simple past tense but instead used simple present tense or another verb tense were included in the count. Second, the number of simple past tense errors was counted. Third, the ratio of error free usage of the simple past tense was

calculated by dividing the number of errors by the number of total uses of the simple past tense. This process was repeated for the past continuous and past perfect tenses.

**Oral Fluency Analysis**

The oral fluency analysis was focused on self-repairs and speech rate. Self-repairs were made up of repetitions and self-corrections (restarts, rephrasing, partial words, and reconstructions). Repetitions were coded where previous versions were placed into parentheses {…} (e.g., {*She she*} *she was watching a program on TV one day about sea animals*) (Foster et al., 2000). Both repeated single words (e.g., one repetition, {*she*} *she*, or two repetitions, {*she, she*} *she*), and repeated phrases (e.g., one repetition, {*she is going to school*} *she is going to school*, or two repetitions, {*she is going to school, she is going to school*} *she is going to school*) were counted as one repetition, regardless of the number of times the word was repeated or the number of words repeated in the repeated phrase. The number of repetitions was calculated by counting the number of coded repeated words and phrases together. For self-corrections, restated, rephrased, partially produced, and reconstructed words or phrases were coded with an em dash (—) at the end of the reconstructed or false starts to indicate that they were incomplete words or phrases. Both self-corrected single words (e.g., she—he) and self-corrected phrases (e.g., she is watching TV—she was watching), regardless of the number of words or the number of times they were corrected, were counted as one self-correction. The number of self-corrections was calculated by counting the number of em dashes (—) coded. Finally, the number of self-repairs was calculated by adding the total number of repetitions and self-

corrections together. For the speech rate analysis, self-repairs were pruned before the number of syllables was calculated. The total number of syllables produced in each pruned transcription by each participant was counted using the following website: http://www.wordcalc.com (Wordcalc, 2019). Wordcalc is free online syllable counting software. Speech rate was measured by dividing the total number of syllables by narration time. Although percentage statistics might vary depending on the length of text produced by individual participants, this was controlled by asking participants to speak as much as possible within their allocated two-minutes. Finally, breakdown fluency was not measured in the present study because I found in the pilot study that few pauses were identified in the narrative task because the participants were instructed to speak as much as possible during the two-minutes narration.

**Weekly Analysis of Nine Participants**

To measure the trajectory of changes in syntactic complexity, syntactic accuracy, and oral fluency over time from the effects of form-focused instruction, weekly narrations of three participants from each group were analyzed (three participants x three groups = nine participants). In order to identify three representative participants from each group, three average achievers were selected based on their speaking proficiency improvement at the end of the research period. Analyzing the performance of average achievers excluded potential issues with outliners as top and bottom achievers are likely to develop speaking proficiency differently from average achievers. The improvement in speaking proficiency of each participant in each group was calculated by adding the gain

score (posttest – pretest) of the six CAF measurements (mean length of T-unit, clause/T-unit ratio, percentage of error-free T-units, percentage of accurate past tense verb usage, speech rate, and self-repair) into one composite *z*-score. Among each group, all participants were ranked on a scale according to their composite *z*-score from the largest to the smallest. This process was repeated for all three groups and the nine selected participants were given pseudonyms (Table 6).

Table 6. *Average Composite z-Score Achievers of the Three Groups*

| Participant | Comparison | FFI | FFI + PF |
|---|---|---|---|
| 1 | Yuka (0.24) | Daiki (0.88) | Mana (0.13) |
| 2 | Yuu (0.20) | Daiichi (0.45) | Manabu (-0.25) |
| 3 | Yusuke (-0.42) | Daisuke (-0.69) | Musashi (-0.67) |

*Note*. Pseudonyms are used. Gain composite *z*-score is inside parentheses. FFI = Form-focused instruction; FFI + PF = Form-focused instruction and peer feedback.

The comparison group consisted of Yuka, Yuu, and Yusuke; The FFI group consisted of Daiki, Daiichi, and Daisuke, and; the FFI + PF group consisted of Mana, Manabu, and Musashi. Among the nine participants, Yuka was one of the weakest English speakers and produced mostly short sentences prior to the start of the study. Yusuke produced a large amount of content in his narrations. Daisuke was one of higher-proficiency English speakers and displayed accurate use of verb tenses prior to the research. Daiki stuttered severely when speaking English. He also produced a large number of Japanese filler pauses such as *eee*, which is similar to English filler pause *um*. Daiichi was the weakest English speaker out of the nine participants. Mana was a skilled English speaker who spoke with confidence and produced almost no filler pauses. Manabu was the highest-proficiency English speaker and also the top achieving student

in his English class. He endeavored to improve his narrations by using complex vocabulary and structures, and narrated the largest amount of content. Manabu produced a large number of self-corrections and filler pauses during his attempts to improve his narrations. Musashi was a proficient speaker who demonstrated high grammatical accuracy prior to the start of the research.

Seven recorded cartoon narrations between the pretest and the posttest, one narration per week by the nine participants were orthographically transcribed, pruned, and analyzed using the abovementioned steps. Numerical changes of the six CAF measurements of all nine representative participants were analyzed to identify trends in the changes within and between the three groups. Finally, weekly qualitative changes in each individual participant were analyzed.

**The Rasch Model for Measurement**

Data collected from the questionnaires were imported into WINSTEPS version 3.64.2 (Linacre & Wright, 2007) and analyzed using the Rasch model to create interval measures. Rasch analysis of raw data is a prerequisite before the analysis of the questionnaire results can occur. The first Rasch analysis concerned the English Speaking Learner Affect Questionnaire, which was designed to measure three latent constructs, classroom English speaking anxiety, English speaking self-efficacy, and desire to speak English. The second Rasch analysis was focused on the Self-Assessment of Speaking Task Questionnaire, which was designed to measure three latent constructs; the participants' self-assessed syntactic complexity, self-assessed syntactic accuracy, and

self-assessed oral fluency of their narrative performances. The six-point Likert scale was used for both questionnaires. Linacre (2002a) suggested three criteria for well-constructed Likert scale questionnaires: there should be at least 10 observations for each category on the scale, the outfit MNSQ statistics for each step should be less than 2.0, and separation between categories should be no less than 0.59 logits for a six-point scale. Item difficulty estimates, person ability estimates, and the fit of items and persons to the Rasch model were investigated to examine the reliability of the data. Finally, Rasch principal component analyses of item residuals were conducted to assess the dimensionality of the target latent constructs.

In addition, a multifaceted Rasch analysis was used to provide validation evidence for the ratings provided by the human raters, who evaluated the participants' speaking proficiency in terms of syntactic complexity, syntactic accuracy, and oral fluency. This analysis was conducted using FACETS version 3.71.4 (Linacre, 2013).

**Principles of the Rasch model.** Rasch analyses can be used to examine the quality of many types of data to provide validation evidence and a framework where test developers and researchers can compare their data across studies (Bond & Fox, 2007). In the Rasch model, raw questionnaire data is transformed into logits and the logits for both the item difficulty estimates and person ability estimates are placed on the same scale. Person ability is calculated based on the total number of correct responses by calculating each person's responses to the items and item difficulty is calculated from the total number of responses by the respondents (Bond & Fox, 2007). The Rasch model therefore

measures the relationship between item difficulty and person ability. Rasch analysis assumes that items differ in difficulty and people differ in ability; therefore, a higher ability person has a greater probability of correctly responding to an item than a lower ability person, and more difficult items will be answered correctly by fewer people compared to easier items (Rasch, 1960).

The Rasch model can also evaluate the degree to which responses to items can produce a single, coherent construct. Therefore, it can be used to confirm the dimensionality of latent constructs measured by Likert-scale questionnaires. In the Rasch model, unidimensional interval level measurement scales can be created using principal component analysis of item residuals when person ability and item difficulty are preserved as response probabilities (Bond & Fox, 2007). More details of the principal component analysis of item residuals are provided below.

However, there are also potential concerns with using the Rasch model. In order for the Rasch model to produce true interval results, the data must be infinite and the underlying assumptions must be met in terms of item and person fit statistics, reliability estimates, the separation index, and unidimensionality. However, while it is ideal to meet all above criteria, in reality, items and persons who do not meet the fit criteria need to be deleted from the model. The deletion of data would happen at the expense of decreasing the sample population.

**Rasch fit statistics.** Rasch fit statistics indicate how well items and participants fit the Rasch model; misfitting participants and items are often excluded from analyses,

as they fail to meet the assumptions of fundamental measurement (Linacre, 2002a). Misfitting persons might not have answered items seriously, they might have guessed the answer, or they might have been careless when answering. Misfitting items might have been poorly written, they might not assess the same construct as other items, or they might be redundant.

Two types of Rasch fit statistics, mean square (MNSQ) and standardized (ZSTD) were calculated. Infit and outfit statistics were calculated for both of them. Outfit MNSQ is the average of the standardized residual variance among a group of items or persons. The outfit MNSQ statistic indicates the degree of unexpected responses within a group of test-takers or a set of items. If persons or items fit the model perfectly, the outfit MNSQ and infit NSQ statistics will be 1.00. Infit MNSQ statistics higher than 1.00 indicate a higher proportion of inconsistent responses than other items in the scale. Infit MNSQ statistics lower than 1.00 indicate a lower proportion of inconsistent responses than the other items in the scale. Outfit MNSQ is similar to infit MNSQ except that it is more sensitive to outliers. Outfit MNSQ statistics higher than 1.00 indicate a higher than expected proportion of erratic responses. Outfit MNSQ statistics lower than 1.00 indicate that there are fewer such responses than would be expected under normal model assumptions. Outfit MNSQ is calculated using the following formula:

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

where $\boldsymbol{Z_{ni}}^{\mathbf{2}}$ is the squared standard residuals, which is summed to form a chi-square statistic divided by N. Both infit MNSQ and outfit MNSQ indices range from 0 to

infinity. A high outfit MNSQ value for an item occurs when the item was answered unexpectedly by participants, for example, incorrectly answered by participants whose ability is higher than the item's difficulty estimate and/or when the item is answered correctly by participants whose ability is lower than the item's difficulty estimate.

Infit MNSQ, which is the squared residuals divided by their expectation, gives more weight to the performance of persons or items with ability measures closer to the parameter values, whereas outfit MNSQ is more sensitive to the influence of outlying scores (Fan & Bond, 2019). Infit MNSQ is calculated using the following formula:

$$\text{infit MNSQ} = \frac{\Sigma Z_{ni}^2 Wni}{\Sigma Wni}$$

where $\mathbf{Z_{ni}}^{\mathbf{2}}$ is the squared standardized residual and $\mathbf{W_{ni}}$ is the individual variances. Weight was placed on the infit MNSQ statistics in this study because outfit MNSQ statistics are more sensitive to the influence of outliners (Bond & Fox, 2007). Linacre (2002b) suggested that values larger than 2.0 indicate that there is more unexplained noise than explained noise, which indicates that there is more misinformation than information observed. According to Fisher's (2007) benchmarks, statistics smaller than 0.33 and larger than 3.00 are poor, those between 0.34 to 2.90 are fair, values between 0.50 to 2.00 are good, those between 0.71 to 1.40 are very good, and statistics between 0.77 to 1.30 are excellent. In this study, infit MNSQ and outfit MNSQ statistics between 0.50 and 1.50 were considered as indicators of acceptable fit to the Rasch model (Linacre, 2007).

Standardized fit statistics: outfit ZSTD and infit ZSTD were also calculated because infit and outfit mean square statistics do not consider N-size and thus a larger of number of participants would lead to values closer to 1 for both statistics. Standardized infit and outfit statistics (usually reported as t or z) are sensitive to sample size. Outfit ZSTD is the standardized *z*-score of the outfit statistic, used to assess item fit. The ZSTD statistics reflect the probability that the misfit of the outfit statistics observed is statistically significant. A Wilson-Hilferty transformation of the outfit MNSQ statistic produces the outfit ZSTD statistic, which corresponds to a *t*-statistic with infinite degrees of freedom. Outfit ZSTD tests the hypothesis that data fit the Rasch model exactly. Infit ZSTD statistics reflect the probability that the infit ZSTD statistic is statistically significant. Bond and Fox (2007) suggested acceptable ranges for standardized statistics to be greater than –2.00 and less than +2.00 because infit ZSTD and outfit ZSTD values higher than +2.00 or lower than –2.00 correspond to a Type 1 error rate of .05. Outfit ZSTD is calculated using the following formula:

outfit ZSTD $= \left(\boldsymbol{v_i^{1/3} - 1}\right))\mathbf{3/q}_i) + (\mathbf{q}_i\mathbf{/3})$.

The outfit ZSTD statistic is the result of a transformation applied to the outfit MNSQ statistic. The transformation renders it as an approximate t-statistic.

**Rasch reliability estimates.** In addition to fit estimates, Rasch item and person reliability estimates were calculated. Rasch item and person reliability, which is analogous to Cronbach's alpha, is calculated as follows:

$$R_I = \frac{SA_I^2}{SD_I^2}$$

where the denominator $\boldsymbol{SD_I^2}$ shows the total person variability, which indicates how much people differ on the measure of interest (Bond & Fox, 2007). The numerator $\boldsymbol{SA_I^2}$ represents the reproducible ability of the variability (i.e., the amount of variance that can be reproduced by the Rasch model). Reliability estimates provide further validity evidence for the questionnaires, as they indicate how accurately the person and item placements can be replicated when the same instruments are administrated to a different group of participants (with similar ability) as well as when the same group of participants is given a different set of items (with similar difficulty level). Rasch person and item reliability indices range between 0 and 1 (Bond & Fox, 2007). In this study, .81 was used as the benchmark for adequate item and person reliability estimates (Fisher, 2007). According to Fisher (2007), estimates < .67 are poor, .67 to .80 are fair, .81 to .90 are good, .91 to .94 are very good, and estimates > .94 are excellent. In addition to reliability estimates, standard error statistics for items and persons were also calculated.

**Rasch separation index.** The Rasch separation index is an estimate of the true measures of standard deviation divided by standard error statistics (Fisher, 1992). Separation estimates indicate the spread of items and persons on the measured variable continuum where a higher value indicates more reliability in defining a hierarchy of items and persons. The Rasch item separation is calculated as follows:

$$G_l = \frac{SA_l}{SE_l}$$

where $\boldsymbol{SE_i}$ is the root mean square measurement (RMSE). Compared to reliability estimates, the Rasch person separation index is a more sensitive measure of how well items differentiate the participants for a given variable. Values range from zero to infinity, with higher values indicating that the measure better distinguishes between the respondents. Values greater than 2.00 are regarded as providing acceptable separation (Fisher, 2007). Measures with low person separation do not measure differences in a sample and therefore are of limited use. Low person separation statistics with a moderate population sample size suggests that the instrument is not sensitive enough to distinguish between higher- and lower-proficiency participants and therefore more items might be needed. Moreover, low item separation statistics suggest that the population sample is not large enough to adequately define the item difficulty hierarchy.

**Rasch principal component analysis of item residuals.** A Rasch principal components analysis (PCA) of item residuals was used to investigate the dimensionality of the questionnaire data and confirm the hypothesized constructs. The Rasch PCA of item residuals identifies systematic unexplained variance in a data set after the fit statistics, separation index, and reliabilities are checked. In the Rasch PCA analysis, data obtained from a multidimensional construct are further examined by checking correlations between item residuals. Item residuals that correlate highly (e.g., above a factor loading of .40) are considered to include a significant amount of variance unexplained by the Rasch model and they therefore need to be reexamined to determine construct dimensionality (Bond & Fox, 2007). A measure is considered acceptably

unidimensional if more than 50.00% of the variance is explained by the Rasch measures, while the unexplained variance is less than 10.00%, or the eigenvalue of the unexplained variance is < 3.00 (Linacre, 2007).

**Multifaceted Rasch analysis.** The Multifaceted Rasch Model is an extension of the conventional Rasch model. FACETS (Version 3.57) was used to analyze the data collected from human raters' analytical ratings of the participants' oral performances from the recorded pretest and posttest. FACETS can analyze data that have more than two facets. Examples of facets are raters, test-takers, and rating criteria. FACETS operates on the assumption that attributes can be measured unidimensionally. In addition to being sample and instrument independent, FACETS is also rater independent, which means that different raters with different severities can rate the same sample with the same instrument and produce the same results. In addition, FACETS can successfully analyze data without having all raters assess all examinees. Therefore, in this study, each of the three raters rated approximately 40.00% of the total sample with an overlap of approximately 20.00% (Bond & Fox, 2007). In addition, the multifaceted Rasch model calculates fit statistics, the Rasch separation index, and Rasch reliability estimates. From the fit estimates, the model identifies raters who gave lows score to highly proficient candidates and high scores to low-proficiency candidates. The fit statistics ensure within-rater consistency, indicating whether raters are consistent in their own ratings.

## Analysis

The following analyses were used to answer the research hypotheses. Hypothesis 1a stated that the participants who receive form-focused instruction (FFI group) and those who receive form-focused instruction and peer feedback (FFI + PF) will improve syntactic accuracy because form-focused instruction explicitly directs them to pay attention to linguistic form (DeKeyser, 1998; Goh & Burns, 2012; Norris & Ortega, 2000; Spada & Tomita, 2010). The comparison group participants, who will not receive form-focused instruction, will not improve syntactic accuracy because they are not explicitly directed to pay attention to the target forms.

Hypothesis 1b stated that all three groups will improve syntactic complexity and oral fluency because task repetition has been found to lead to such improvements (Ahmadian, 2011; Bygate, 2001). In order to examine whether the participants improve syntactic complexity, syntactic accuracy, and oral fluency, and to what extent they improve after participating in the intervention, within-group analyses are conducted using one-way repeated-measures ANOVAs with the pretest and posttest scores. One-way repeated-measures ANOVAs were chosen over paired-sample *t*-tests because they compute the effect size statistic, $\eta^2$. Six one-way repeated-measures ANOVAs for each six of the CAF measurements were conducted for the FFI group (See Table 6). This process was repeated for the FFI + PF and comparison groups, respectively.

Hypothesis 2a stated that the FFI group will improve syntactic accuracy more than the comparison group because form-focused instruction explicitly directs them to pay attention to form, especially the use of past tense forms. Task repetition alone does

not explicitly direct comparison group participants to pay attention to the target forms and to notice errors in their use of the target forms. The FFI + PF group will improve syntactic accuracy more than the FFI group because more explicit form-focused instruction leads to more syntactic accuracy development (DeKeyser, 1998; Goh & Burns, 2012; Norris & Ortega, 2000; Spada & Tomita, 2010).

Hypothesis 2b states that there will not be significant differences in syntactic complexity nor oral fluency between the FFI + PF, FFI, and the comparison groups because the form-focused instruction provided in this study does not target syntactic complexity nor oral fluency development; all groups might improve as a result of the effects of task repetition (Ahmadian, 2011; Bygate, 2001).

To examine if the FFI + PF group improves more in syntactic accuracy than the FFI group, and if the FFI group improves more than the comparison group after seven weeks of intervention, a one-way MANOVA was conducted. The independent variable was group (three levels: comparison, FFI, and FFI + PF) and the dependent variables were the gain scores of six CAF measurements (posttest - pretest score). Pearson correlations were calculated among the six CAF variables on the pretest to determine that they were no high or low correlations as they are considered to be unsuitable for MANOVA analyses (Field, 2009). Following the MANOVA, one-way ANOVAs were conducted for each of the six CAF measures. The independent variable was group and dependent variable was the gain score for each measure.

In addition to analyses of the six CAF measurements, human raters rated the performances of the three groups. The ratings were analyzed with the FACETS software

and fair average measures were produced. Gain scores (posttest - pretest score) were calculated. A one-way ANOVA was conducted where the independent variable was group and dependent variable was the gain score for the fair average measure.

Hypothesis 3 stated that the FFI + PF group will improve syntactic accuracy more quickly week by week than the FFI group and the comparison group because the FFI + PF group receives more form-focused instruction and form-focused instruction increases syntactic accuracy (Ahmadian & Tavakoli, 2011; Mochizuki & Ortega, 2008; Williams, 2005; Yuan & Ellis, 2003). To measure the changes in syntactic complexity, syntactic accuracy, and oral fluency over time from the effects of form-focused instruction, weekly narrations of three representative participants from each group ($n$ = 9) were transcribed and analyzed. Week to week changes in the six CAF measurements—mean length of T-unit, clauses/T-unit ratio, percentage of error-free T-units, percentage of accurate past tense verb usage, speech rate, and self-repair—were analyzed to identity possible patterns of change in each group and in each of the nine participants.

Hypothesis 4a stated that the FFI + PF group will increase in classroom English speaking anxiety more than the FFI and comparison groups, and the FFI group will increase more than the comparison group because the FFI + PF group receives more form-focused instruction (Saito, 2004; Trang et al., 2013).

Hypothesis 4b stated that the FFI + PF group will increase speaking self-efficacy more than the FFI and comparison groups, and the FFI group will increase more than the comparison group because form-focused instruction can increase self-confidence (Van Batenburg et al., 2019).

Hypothesis 4c stated that the FFI + PF group will decrease in desire to speak English more than the FFI and comparison groups, and the FFI group will decrease more than the comparison group because form-focused instruction is likely to decrease learners' willingness to communicate (Van Batenburg et al., 2019).

To measure the effects of form-focused instruction on learner affect, the English Speaking Learner Affect Questionnaire was administered after the participants completed the pretest and posttest speaking narration. Rasch person ability estimates were used to examine changes in classroom English speaking anxiety, English speaking self-efficacy, and desire to speak English. Three one-way ANCOVAs were conducted to examine possible differences between the three groups in classroom English speaking anxiety, English speaking self-efficacy, and desire to speak English respectively. The first ANCOVA was conducted to examine whether (a) the FFI + PF group increases classroom English speaking anxiety more than the FFI group and (b) the FFI group increases more than the comparison group. The second ANCOVA was conducted to examine whether (c) the FFI + PF group increases more English speaking self-efficacy than the FFI group and (e) the FFI group increases more than the comparison group. The third ANCOVA was conducted to examine whether (f) the FFI + PF group decreases desire to speak English more than the FFI group, and (g) the FFI group decreases more than the comparison group after seven weeks of intervention. The independent variable was group (three levels: comparison, FFI, and FFI + PF) and the dependent variables were the three affective variables (classroom English speaking anxiety, English speaking

self-efficacy, and desire to speak English) at posttest. The covariate was the three affective variables at pretest.

Hypothesis 5a stated that the FFI + PF group will have higher self-assessments of syntactic accuracy than the FFI and comparison groups. The FFI group will have higher self-assessments of syntactic accuracy than the comparison group because more form-focused instruction can lead to greater syntactic accuracy gains (DeKeyser, 1998; Goh & Burns, 2012; Norris & Ortega, 2000; Spada & Tomita, 2010) and therefore higher self-assessments (Ma & Winke, 2019; Malabonga et al., 2005; Poehner, 2012).

Hypothesis 5b stated that the FFI + PF group and the FFI group will not have higher self-assessments for syntactic complexity or oral fluency than the comparison group because form-focused instruction does not target improvements in syntactic complexity or oral fluency.

To measure learners' assessment of their own speaking performances, the Self-Assessment of Speaking Task Questionnaire was administered after the participants completed the pretest and posttest narrations. Rasch person ability estimates were used to examine changes in self-assessed syntactic complexity, self-assessed syntactic accuracy, and self-assessed oral fluency. A one-way MANOVA was conducted to examine whether (a) the FFI + PF group self-assesses their syntactic complexity higher than the FFI group, (b) the FFI group self-assesses themselves higher than the comparison group, (c) the FFI + PF group self-assesses their own syntactic accuracy higher than the FFI group, (d) the FFI group self-assesses themselves higher than the comparison group, (e) the FFI + PF group self-assesses their oral fluency higher than the FFI group, and (f) the FFI group

self-assesses themselves higher than the comparison group after seven weeks of intervention. The independent variable was group (three levels: comparison, FFI, and FFI + PF) and the dependent variables were the gain scores of three self-assessed construct (posttest - pretest score). Following the MANOVA, three one-way ANOVAs were conducted for each of the three self-assessed constructs. The independent variable was group (three levels: comparison, FFI, and FFI + PF) and the dependent variables were the gain (posttest - pretest logit) in self-assessed syntactic complexity, self-assessed syntactic accuracy, and self-assessed oral fluency.

## Summary

In this chapter I described the research setting, participants, selection of target forms for intervention, the two intervention groups and the comparison group, as well as the instruments. The chapter concluded with a description of the procedures that were employed, the data that were deleted and pruned, the analyses, and a description of how the hypotheses are analyzed.

In the next chapter, I present the results of the preliminary analyses to examine the dimensionality and validity evidence of six measures. I then present the multifaceted Rasch analysis for the analytical ratings, discuss the three types of past tense forms observed in the present study, and discuss the focus on the simple past tense.

# CHAPTER 4

# PRELIMINARY ANALYSES

In this chapter I first present the results of the Rasch analyses for the English Speaking Learner Affect Questionnaire and Self-Assessment of Speaking Task Questionnaire. Second, I present the results of the multifaceted Rasch analysis for the ratings of the spoken performances. Finally, I describe the frequency of the use of the simple past, past continuous, and past perfect tenses observed in the narrations.

## Preliminary Analyses of English Speaking Learner Affect Questionnaire

The English Speaking Learner Affect Questionnaire was administered after the participants completed the pretest and posttest narrations in order to measure changes in classroom speaking anxiety, English speaking self-efficacy, and desire to speak English. A Rasch analysis was conducted for each of the three latent constructs to examine their item fit to the Rasch model, construct dimensionality, reliability, and to produce interval measures. Although the same questionnaire items were administered on both occasions, item difficulty estimates and the rating scale can change during different administrations; therefore, a decision was made to stack the pretest and posttest data (Wright, 1996, 2003). Stacking ensures that the same item difficulty estimates and rating scale are used at both measurement occasions; thus, any differences in person ability estimates are due to changes in the participants' responses to the questionnaire items. Separate preliminary

analyses of the pretest and posttest data conducted before stacking the data showed that the questionnaire items displayed acceptable fit to the Rasch model.

**Rasch Analysis of Item Fit and PCA of Item Residuals for Classroom English Speaking Anxiety**

The 16 items measuring Classroom English Speaking Anxiety were analyzed using WINSTEPS version 3.64.2 (Linacre & Wright, 2007). Pretest and posttest data were stacked into a single data set (Wright, 1996); each person was considered a unique individual at each administration; therefore, data from 208 participants were analyzed (i.e., 104 participants x 2 administrations = 208).

The functioning of the six-point Likert-scale (1 = *Strongly disagree*, 2 = *Disagree*, 3 = *Slightly disagree*, 4 = *Slightly agree*, 5 = *Agree*, and 6 = *Strongly agree*) is presented in Table 7. First, the minimum of 10 observations per category was met. Category 6 (*Strongly agree*) had the smallest number of observations, at 178 (5.00%). Second, the outfit MNSQ statistics for all six categories were below 2.00. Third, the separation between adjacent categories was greater than the required 0.59 logits for a six-point scale (Linacre, 2002b). Based on the above results, it was concluded that the six-point Likert-scale performed satisfactorily.

Table 7. *Category Structure Functioning for Classroom English Speaking Anxiety*

| Category | Count | (%) | Infit MNSQ | Outfit MNSQ | Structure measure | Category measure |
|---|---|---|---|---|---|---|
| 1 Strongly disagree | 392 | 12 | 1.19 | 1.19 | None | (-3.22) |
| 2 Disagree | 597 | 18 | 0.91 | 0.89 | -1.86 | -1.66 |
| 3 Slightly disagree | 845 | 25 | 0.87 | 0.93 | -1.07 | -0.57 |
| 4 Slightly agree | 841 | 25 | 0.91 | 0.95 | -0.12 | 0.47 |
| 5 Agree | 478 | 14 | 1.10 | 1.24 | 1.00 | 1.68 |
| 6 Strongly agree | 170 | 5 | 0.96 | 1.02 | 2.06 | (3.36) |

All items except item A13, *I do not ask questions in Japanese*, met the 0.50 to 1.50 criterion (Linacre, 2002a). Item A13 did not explicitly target anxiety towards speaking English, so it was removed from the analysis. In order to check the effect of deleting item A13, the Rasch person measures with and without item A13 were exported into SPSS and a bivariate correlation analysis was run. The Pearson correlation of .99 ($p < .01$) indicated that deleting A13 did not affect the person measures. Table 8 shows the Rasch item statistics for new measure after deleting item A13.

Table 8. *Rasch Item Statistics for Classroom English Speaking Anxiety*

| Item | Measure | *SE* | Infit MNSQ | Infit ZSTD | Outfit MNSQ | Outfit ZSTD | Pt-measure correlation |
|---|---|---|---|---|---|---|---|
| A16 | 1.48 | 0.09 | 1.00 | 0.00 | 0.98 | -0.20 | .59 |
| A14 | 1.28 | 0.08 | 1.16 | 1.60 | 1.11 | 1.10 | .57 |
| A09 | 1.24 | 0.08 | 1.10 | 1.00 | 1.05 | 0.50 | .61 |
| A04 | 0.35 | 0.08 | 0.86 | -1.50 | 0.85 | -1.60 | .74 |
| A05 | 0.22 | 0.08 | 0.86 | -1.50 | 0.86 | -1.50 | .74 |
| A03 | 0.06 | 0.08 | 0.66 | -4.00 | 0.67 | -3.90 | .80 |
| A12 | -0.23 | 0.08 | 0.65 | -4.10 | 0.65 | -4.20 | .82 |
| A06 | -0.24 | 0.08 | 1.23 | 2.30 | 1.43 | 4.00 | .65 |
| A15 | -0.25 | 0.08 | 1.19 | 2.00 | 1.18 | 1.80 | .69 |
| A07 | -0.43 | 0.08 | 0.85 | -1.60 | 0.87 | -1.40 | .74 |
| A10 | -0.47 | 0.08 | 0.66 | -4.00 | 0.65 | -4.20 | .79 |
| A02 | -0.50 | 0.08 | 0.92 | -0.80 | 0.90 | -1.00 | .78 |
| A11 | -0.78 | 0.08 | 1.40 | 3.80 | 1.45 | 4.10 | .56 |
| A08 | -0.84 | 0.08 | 1.03 | -0.30 | 1.03 | 0.40 | .70 |
| A01 | -0.89 | 0.08 | 1.45 | 4.20 | 1.45 | 4.20 | .62 |

The Wright map shown in Figure 4 presents the participants on the left side of the figure. More anxious students are toward the top of the figure and less anxious students are toward the bottom. The items are displayed on the right according to their difficulty; more difficult items are placed toward the top of the figure and less difficult items are placed toward the bottom. Difficult to endorse items means that the respondents did not feel anxious about those items. The mean person ability measure for the persons of -0.31 was close to the mean item difficulty estimate of 0; this result indicated that the items matched the participants well, albeit they were slightly difficult to endorse for this group. Figure 4 also shows that the respondents are well distributed around the item mean.

The most difficult to endorse item (i.e., the item the participants felt least anxious about) was item A16, *I become anxious whether my classmates will laugh when I speak English in group discussions*. The second most difficult item to endorse was item A14, *I become embarrassed when answering the teacher's question in English*, and third most difficult to endorse item was item A09, *I become anxious whether my classmates will laugh when I speak English*. On the other end of the scale, the easiest to endorse item was item A08, *I become anxious when I cannot express my opinion in English*. This response is reasonable, as most people feel anxious when they cannot say what they want to say. Figure 4 indicates that the participants did not feel anxious about speaking English; however, they were worried that other students were better or that other students did not understand them. Therefore, speaking English did not make the participants overly anxious in general, but there was a clear element of social anxiety.

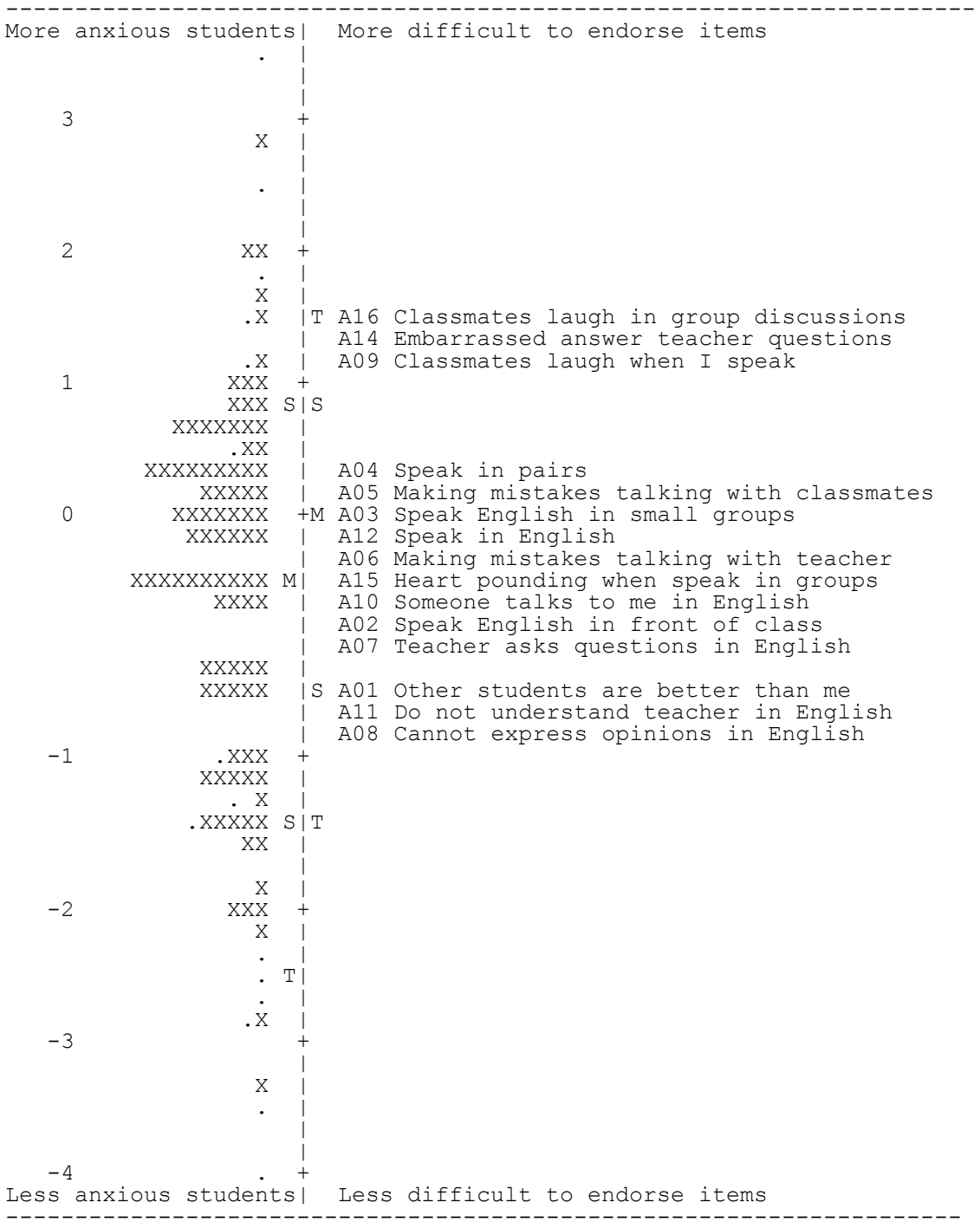



*Figure 4*. Wright map for classroom English speaking anxiety. M = Mean; S = 1 SD; T = 2SDs. Each X = 2 participants.

A Rasch PCA of item residuals analysis was conducted to investigate the dimensionality of the Classroom English Speaking Anxiety items. The Rasch model explained 59.30% of the variance (eigenvalue = 23.30); this percentage exceeded the 50.00% criterion proposed by Linacre (2007). In addition, 35.20% of the variance (eigenvalue = 13.80) was explained by the items, and 7.60% of the variance (eigenvalue = 3.00) was accounted for by the first residual contract. This figure was below the 10.00% criterion, and also met the eigenvalue criterion of < 3.00. The results of the Rasch PCA of item residuals analysis indicated that the Classroom English Speaking Anxiety measure was fundamentally unidimensional.

Table 9 shows the loadings of the remaining 15 items from the Classroom English Speaking Anxiety measure. Items with the highest positive loadings were item A09, *I am afraid my classmates will laugh when I speak English*, item A16, *I become anxious whether my classmates will laugh when I speak English in group discussions*, and item A14, *I become embarrassed when answering the teacher's question in English*. While items A16 and A09 were directly related to learners' fear of not speaking well, item A14 was also related closely to learners' fear of not speaking well as they are afraid of being judged negatively by the teacher when answering questions. These strong positive loading items represent the facet of fear of underachievement. Items with the highest negative loadings were item A02, *I become anxious when I speak English in front of the entire class*, item A03, *I become anxious when I speak English in small groups*, and item A12, *I become anxious when I speak in English*. These three items relate to learners'

discomfort when speaking in English as English is a foreign language. The strong negatively loading items represent the facet of uneasiness in speaking English.

Table 9. *Rasch PCA of Item Residuals for Classroom English Speaking Anxiety*

| Item | Loading | Measure | Infit MNSQ | Outfit MNSQ |
|---|---|---|---|---|
| A09 Classmates laugh at me | 0.66 | 1.24 | 1.10 | 1.05 |
| A16 Classmates laugh in groups | 0.61 | 1.48 | 1.00 | 0.98 |
| A14 Answer teacher's question | 0.54 | 1.28 | 1.16 | 1.11 |
| A06 Make mistakes with teacher | 0.28 | -0.24 | 1.23 | 1.43 |
| A11 Don't understand teacher | 0.16 | -0.78 | 1.40 | 1.45 |
| A07 Asked question by teacher | 0.06 | -0.43 | 0.85 | 0.87 |
| A08 Cannot express opinions | 0.04 | -0.84 | 1.03 | 1.03 |
| A05 Make mistake with classmates | 0.02 | 0.22 | 0.86 | 0.86 |
| A02 Speak English front of class | -0.58 | -0.50 | 0.92 | 0.90 |
| A03 Speak English in small groups | -0.54 | 0.06 | 0.66 | 0.67 |
| A12 Speak in English | -0.47 | -0.23 | 0.85 | 0.65 |
| A01 Worry other students better | -0.37 | -0.89 | 1.45 | 1.45 |
| A04 Speak English in pairs | -0.34 | -0.35 | 0.86 | 0.85 |
| A10 Someone talks English to me | -0.32 | -0.47 | 0.66 | 0.65 |
| A15 Heart pounding speak English | -0.08 | -0.25 | 1.19 | 1.18 |

The Rasch item separation statistic was 8.97 and the person separation statistic was 3.05; item separation was excellent, as it was above the benchmark of 5.0 and person separation was good, as it was above 3.00 (Fisher, 2007). The item reliability estimate was .99, which was excellent, as it was above .94 and person reliability was .90, which was good according to Fisher (2007). As indicated by the separation statistics, the items formed over eight distinct levels of difficulty, while the persons formed approximately three distinct levels.

**Descriptive Statistics for Classroom English Speaking Anxiety**

After conducting the Rasch analysis of the Classroom English Speaking Anxiety measure, the Rasch person ability estimates were exported into SPSS to check the assumption of normality. Initial descriptive statistics showed that normality cannot be assumed because kurtosis was 2.89, which was over the ±2.00 criterion (George & Mallery, 2016). Therefore, two univariate outliers with *z*-score above ±3.29 (Field, 2009) were deleted and assumption of normality was checked again (Case 60 had a *z*-score of 3.58 and Case 63b had a *z*-score of 6.25). The descriptive statistics in Table 10 show that normal distribution can be assumed after deleting the two outliers, as skewness of 0.13 and kurtosis of 0.80 were both under ±2.00.

Table 10. *Descriptive Statistics for Classroom English Speaking Anxiety Items*

| | |
|---|---|
| *M* | -0.26 |
| *SE* | 0.08 |
| 95% CI | [-0.42, -0.11] |
| *SD* | 1.21 |
| Skewness | 0.13 |
| *SES* | 0.17 |
| Kurtosis | 0.80 |
| *SEK* | 0.33 |

*Note*. *N* = 206, SES = Standard error skewness; SEK = Standard error kurtosis.

**Rasch Analysis of Item Fit and PCA of Item Residuals for English Speaking Self-Efficacy**

The 11 items designed to measure English Speaking Self-Efficacy were analyzed using the Rasch rating scale model. Data from the pretest and posttest questionnaire administrations were stacked into a single data set and each person was considered a unique individual at each administration; therefore, data from 208 participants were

analyzed (i.e., 104 participants x 2 administrations = 208). The functioning of the six-point Likert-scale is presented in Table 12. The minimum of 10 observations per category was met, as the smallest number of observations, 67 (3.00%), was found for Category 6 (*Strongly agree*). The outfit MNSQ statistics for all six categories were below the 2.00 criterion. The smallest separation, which was between Category 2 (-2.67) and Category 3 (-1.18), was 1.49. The separation between all categories was greater than the minimum requirement of 0.59 logits for a six-point scale. Based on the above evidence, it was concluded that the performance of the six-point Likert-scale was satisfactory.

Table 11. *Category Structure Functioning for English Speaking Self-Efficacy*

| Category | Count | (%) | Infit MNSQ | Outfit MNSQ | Structure measure | Category measure |
|---|---|---|---|---|---|---|
| 1 Strongly disagree | 101 | 4 | 1.20 | 1.21 | None | (-4.54) |
| 2 Disagree | 284 | 12 | 0.99 | 1.04 | -3.30 | -2.67 |
| 3 Slightly disagree | 573 | 25 | 0.89 | 0.89 | -1.80 | -1.18 |
| 4 Slightly agree | 834 | 37 | 0.93 | 0.93 | -0.63 | 0.48 |
| 5 Agree | 423 | 19 | 0.93 | 0.92 | 1.45 | 2.89 |
| 6 Strongly agree | 67 | 3 | 1.51 | 1.27 | 4.28 | (5.43) |

All of the items except item E01, *I can order food in English in a restaurant*, met the 0.50-1.50 infit MNSQ statistics criterion suggested by Linacre. Item E01 had an infit MNSQ statistic of 1.67, and an outfit MNSQ statistic of 1.68, so it was removed from the analysis. This item might have misfit because it might have been difficult for students who have never travelled outside of Japan to respond to the item. Furthermore, item E01 differs from the other items because it is not directly related to English classroom tasks; it is related to general English skills needed outside the classroom. In order to check the effect of deleting item E01, the Rasch person measures with and without item E01 were

exported into SPSS and a bivariate correlation analysis was run. The Pearson correlation of .99 ($p < .01$) indicated that deleting the item did not substantially affect the person measures. Table 13 displays the individual item statistics for the English Speaking Self-Efficacy items after deleting item E01.

Table 12. *Rasch Item Statistics for English Speaking Self-Efficacy*

| Item | Measure | SE | Infit MNSQ | Infit ZSTD | Outfit MNSQ | Outfit ZSTD | Pt-measure correlation |
|---|---|---|---|---|---|---|---|
| E09 | 1.09 | 0.09 | 0.96 | -0.30 | 0.98 | -0.20 | .75 |
| E10 | 0.88 | 0.09 | 0.72 | -3.20 | 0.76 | -2.60 | .77 |
| E08 | 0.43 | 0.10 | 1.24 | 2.30 | 1.23 | 2.20 | .72 |
| E06 | 0.32 | 0.10 | 0.98 | -0.20 | 1.00 | 0.00 | .76 |
| E11 | 0.20 | 0.10 | 1.30 | 2.80 | 1.33 | 3.10 | .67 |
| E07 | -0.17 | 0.10 | 1.11 | 1.10 | 1.13 | 1.30 | .71 |
| E05 | -0.35 | 0.10 | 0.96 | -0.40 | 0.94 | -0.60 | .76 |
| E04 | -0.48 | 0.10 | 0.71 | -3.20 | 0.71 | 3.20 | .79 |
| E03 | -0.77 | 0.10 | 0.99 | -0.10 | 0.96 | -0.40 | .75 |
| E02 | -1.16 | 0.11 | 0.96 | -0.40 | 0.90 | -1.00 | .76 |

The Wright map shown in Figure 5 indicates that item E10, *I can talk with the teacher in English casually*, was the most difficult item to endorse. Having a conversation in English is difficult and having the conversation with the teacher adds an extra cognitive and affective burden for many students, as it requires not only the necessary speaking proficiency, but often students do not know how to continue a conversation with the teacher. The second most difficult to endorse item was item E09, *I can talk with international students in English casually*, which was slightly easier than item E10 because the respondents probably find it easier to converse with people of a similar age. Both items E09 and E10 were difficult for the participants to endorse because they are considered dialogues. Producing dialogues requires the speakers to take turns,

consider between-turn pauses, allow for interruptions by the other speaker, and continue the simultaneity of talking (Tavakoli, 2016). Dialogue is not only a collaborative activity, but it also requires that learners anticipate what the other speaker is going to say; thus, second language learners whose L2 knowledge is not yet adequately developed might find dialogues difficult because they need considerable L2 knowledge to successfully complete turn-taking (Tavakoli, 2016). The easiest item to endorse was item E02, *I can introduce myself in English*. This item places the least demand on speaking ability, as it is a one-way monologue and most Japanese university students have experience giving self-introductions in English. The left side of Figure 5 shows that most participants are located between +3 to -2 logits, indicating that they are well spread out in terms of their self-efficacy for speaking English. A few outliers are located around +6 and -6 logits, indicating that they have extremely high or low perceptions of their speaking ability.

A Rasch PCA of item residuals analysis was conducted to investigate the dimensionality of the English Speaking Self-Efficacy measure. The Rasch model explained 58.50% of the variance (eigenvalue = 14.20); this percentage exceeded the 50.00% criterion proposed by Linacre (2007). In addition, 22.90% of the variance (eigenvalue = 5.60) was explained by the items. Finally, 10.90% of the unexplained variance (eigenvalue = 2.60) was in the first residual contract. Although 10.90% did not meet the < 10.00% criterion, the eigenvalue criterion of < 3.00 was met, so the results indicated that this set of items is fundamentally unidimensional.

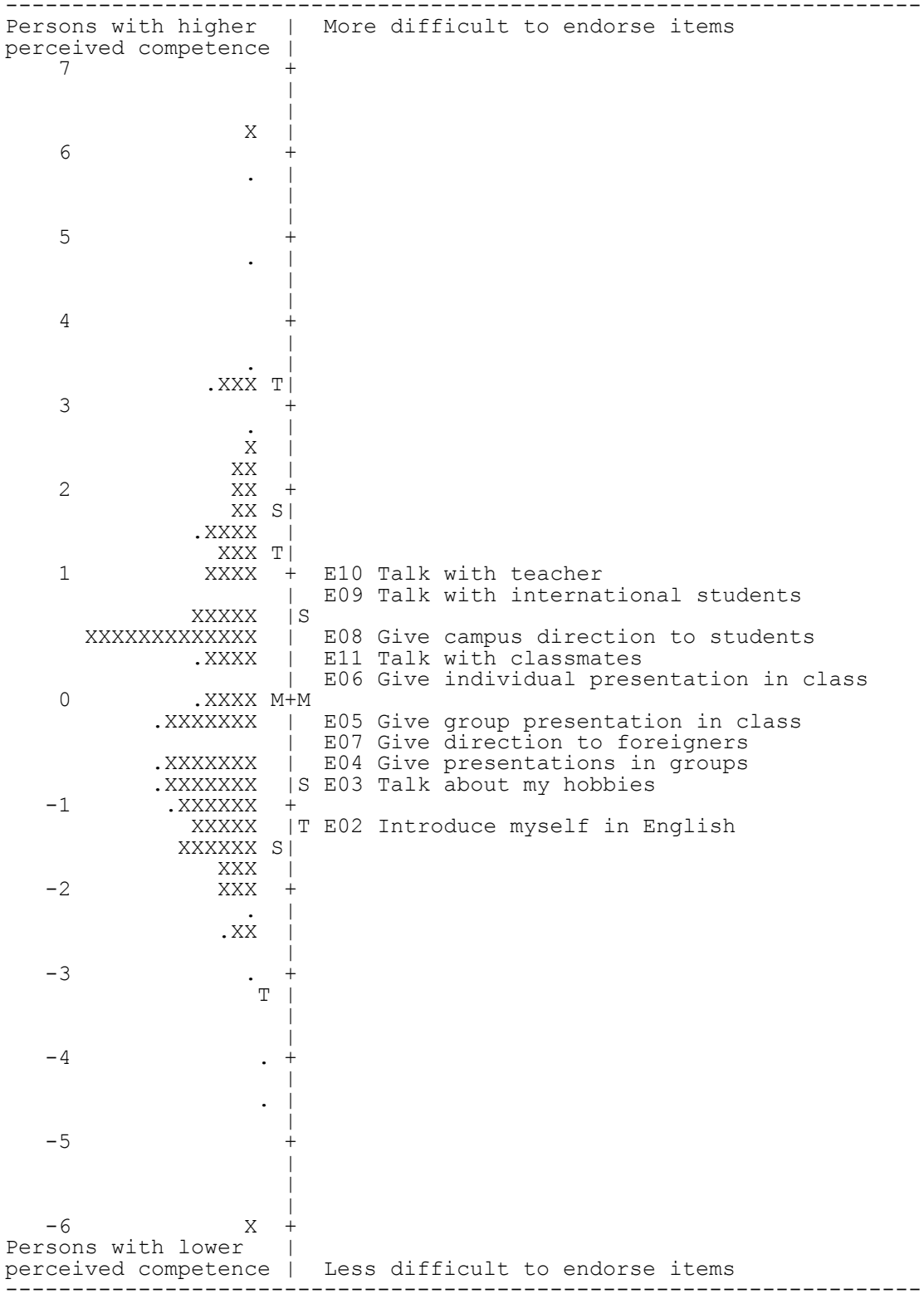



*Figure 5*. Wright map for English speaking self-efficacy. M = Mean; S = 1 SD; T = 2SDs. Each X = 2 participants.

Table 13 shows the Rasch loading of the 10 items of the English Speaking Self-Efficacy measure. High positive loadings occurred with item E06, *I can give an individual English presentation in front of class*, item E05, *I can give a group English presentation in front of the class*, and item E04, *I can give a presentation in English in groups*. These three items are all concerned with giving presentations in English, so they can be considered to belong to a presentation facet. On the other hand, four items had high negative loadings: item E08, *I can give campus directions to international students*, item E09, *I can talk with international students in English casually*, item E07, *I can give street directions in English to foreigners*, and item E10, *can talk with the teacher in English casually*. These four items all concern conversational interactions with a teacher, international students, or people while giving directions. Therefore, these four items relate to a conversational interaction facet. Items in the presentation facet differ from the items in the conversational interaction facet because presentations are monologues. While the construct is considered fundamentally unidimensional, the strong positive and strong negative loadings suggest that two facets might exist within the English Speaking Self-Efficacy construct. Therefore, caution needs to be given to these two possible dimensions when interpreting the results of the English Speaking Self-Efficacy measure.

Rasch item separation was 6.53 and the person separation statistic was 3.06; item separation was excellent, as it was above the benchmark of 5.0 and person separation was good, as it was above 3.0 (Fisher, 2007). The item reliability estimate was .98, which was excellent, as it was above .94, and person reliability was .90, which was good, as it was between .81 and .90 (Fisher, 2007). As indicated by the separation statistics, the ten items

for the English Speaking Self-Efficacy measure formed over six distinct levels of difficulty, while the persons formed approximately three distinct levels.

Table 13. *Rasch PCA of Item Residuals for English Speaking Self-Efficacy*

| Item | Loading | Measure | Infit MNSQ | Outfit MNSQ |
|---|---|---|---|---|
| E06 Give individual presentation | 0.73 | 0.32 | 0.98 | 1.00 |
| E05 Give group presentation | 0.69 | -0.35 | 0.96 | 0.94 |
| E04 Give presentation in groups | 0.67 | -0.48 | 0.71 | 0.71 |
| E02 Introduce myself in English | 0.21 | -1.16 | 0.96 | 0.90 |
| E03 Talk about my hobbies | 0.18 | -0.77 | 0.99 | 0.96 |
| E08 Give campus directions | -0.59 | 0.43 | 1.24 | 1.23 |
| E09 Talk with international students | -0.55 | 1.09 | 0.96 | 0.98 |
| E07 Give direction to foreigners | -0.49 | -0.17 | 1.11 | 1.13 |
| E10 Talk with teacher | -0.42 | 0.88 | 0.72 | 0.76 |
| E11 Talk with classmates | -0.14 | 0.20 | 1.30 | 1.33 |

**Descriptive Statistics for English Speaking Self-Efficacy**

After conducting the Rasch analysis of the English Speaking Self-Efficacy measure, the person ability estimates were exported into SPSS in order to check assumption of normality. Initial descriptive statistics showed that kurtosis was 3.48, which indicated that normality cannot be assumed, as it was over ±2.00 (George & Mallery, 2016). Therefore, two outliers with *z*-scores above ±3.29 (Field, 2009) were deleted and the assumption of normality was checked again. The descriptive statistics in Table 14 show that after deleting the outliers, skewness was 0.90 and kurtosis was 1.90. Therefore, the normality of the distribution was assumed because both statistics were under ±2.00 (George & Mallery, 2016).

Table 14. *Descriptive Statistics for English Speaking Self-Efficacy Items*

| | |
|---|---|
| *M* | 0.08 |
| *SE* | 0.10 |
| 95% CI | [-0.11, 0.29] |
| *SD* | 1.47 |
| Skewness | 0.90 |
| *SES* | 0.17 |
| Kurtosis | 1.90 |
| *SEK* | 0.34 |

*Note*. *N* = 206; SES = Standard error skewness; SEK = Standard error kurtosis.

**Rasch Analysis of Item Fit and PCA of Item Residuals for Desire to Speak English**

The nine items measuring Desire to Speak English were analyzed with the Rasch rating scale model. Data from both the pretest and posttest questionnaire administrations were stacked into a single data set; each person was considered a unique individual at each administration; therefore, data from 208 participants were analyzed (i.e., 104 participants x 2 administrations = 208). The responses to the six-point Likert-scale were examined and presented in Table 15. In the original category function analysis, the outfit MNSQ statistic of Category 1 was 4.35, which exceeded the 2.00 criterion (Linacre, 2002b). The separation between adjacent categories was greater than the minimum criterion of 0.59 (Linacre, 2002b). However, Category 1 did not meet the minimum of 10 observations per category. Therefore, Category 1 and Category 2 were combined, and the analysis was rerun. The new Category 1 (*Disagree*) had 32 observations (2.00%) with an outfit MNSQ statistic of 1.46, which was below the 2.00 criterion (Linacre, 2002b); the outfit MNSQ statistics of the remaining five categories were all below the 2.00 criterion. Finally, the separation between the categories in the new five-point category was greater than 0.81 logits (Wolfe & Smith, 2007), with the smallest separation between Category 1 (-3.69) and Category 2 (-2.17) to be 1.50. Therefore, it was concluded that the

performance of the new five-point Likert-scale for the Desire to Speak English measure was satisfactory and all categories functioned well.

Table 15. *Category Structure Functioning for Desire to Speak English*

| Category | Count | (%) | Infit MNSQ | Outfit MNSQ | Structure measure | Category measure |
|---|---|---|---|---|---|---|
| 1 Combine disagree | 32 | 2 | 1.26 | 1.46 | None | (-3.69) |
| 2 Slightly disagree | 76 | 4 | 1.18 | 1.19 | -2.22 | -2.17 |
| 3 Slightly agree | 459 | 25 | 0.89 | 1.03 | -2.05 | -0.41 |
| 4 Agree | 591 | 32 | 0.95 | 1.02 | 0.99 | 2.15 |
| 5 Strongly agree | 701 | 38 | 0.93 | 0.95 | 3.27 | (4.44) |

All items except item D04, *I want to speak English with other Japanese people*, and item D07, *I want to go to overseas countries,* met the 0.50-1.50 infit MNSQ criterion suggested by Linacre (2002a). Item D04 was deleted because it had an infit MNSQ statistic of 2.04, and an outfit MNSQ statistic of 2.15, which exceeded the 0.50-1.50 criterion. Item D04 can be considered a difficult item, as some students dislike communicating with other students who share the same L1 in the target language. Item D07 had an infit MNSQ statistic of 1.67, and an outfit MNSQ statistic of 1.77, which also exceeded the 0.50-1.50 criterion. The desire to go overseas is perhaps more related to leisure and does not relate directly with the desire to speak English. In order to check the effect of deleting items D04 and D07, the Rasch person measures without items D04 and D07 were exported into SPSS and a Pearson correlation analysis was run. The Pearson correlation of .99 ($p < .01$) indicated that excluding these two items did not substantially affect the Rasch person measures; therefore, items D04 and D07 were deleted. Table 16 shows the individual items for Desire to Speak English after deleting items D04 and D07.

Table 16. *Rasch Item Statistics for Desire to Speak English*

| Item | Measure | *SE* | Infit MNSQ | Infit ZSTD | Outfit MNSQ | Outfit ZSTD | Pt-measure correlation |
|---|---|---|---|---|---|---|---|
| D11 | 0.78 | 0.12 | 1.00 | 0.10 | 1.04 | 0.40 | .81 |
| D09 | 0.71 | 0.12 | 0.95 | -0.50 | 0.91 | -0.80 | .82 |
| D10 | 0.33 | 0.12 | 0.61 | -4.10 | 0.63 | -3.70 | .86 |
| D03 | 0.17 | 0.12 | 0.75 | -2.40 | 0.73 | -2.60 | .83 |
| D08 | 0.08 | 0.12 | 1.06 | 0.60 | 1.09 | 0.80 | .77 |
| D01 | -0.03 | 0.13 | 1.03 | 0.30 | 1.15 | 1.30 | .77 |
| D02 | -0.14 | 0.13 | 0.77 | -2.20 | 0.71 | -2.70 | .82 |
| D06 | -0.17 | 0.13 | 1.42 | 3.40 | 1.51 | 3.70 | .73 |
| D05 | -1.72 | 0.15 | 1.39 | 3.10 | 1.66 | 3.20 | .60 |

The Wright map (Figure 6) presents the participants on the left according to their level of desire to speak English. Students with more self-perceived desire to speak English are located toward the top of the figure and those with less desire to speak English are toward the bottom; more difficult to endorse items are placed toward the top of the figure and less difficult to endorse items are placed toward the bottom. Figure 6 has positive values on the left-hand side where many participants' ability estimates ranged from 0 to +4 logits; thus, most of the participants had a strong desire to speak English. There is a large cluster of participants located at +6 logits indicating the presence of a large group of participants with very strong desire to speak English. On the other hand, there were also a few participants with very low desire to speak English. On the right-hand side, most items clustered around the mean, but item D05, *I want to express my opinions and ideas in English*, is located at the bottom of the difficulty scale level; thus, it was very easy to endorse. The most difficult to endorse items were item D11, *I want to speak with the English teacher in English*, and item D09, *I want to find opportunities to speak English outside class*. Item D11 was the most difficult to endorse item; the participants were university students and some of them did not feel at ease when

talking with their teachers in English. Item D09 was the second most difficult to endorse item as the participants find it difficult to find opportunities to speak English while living in the Japanese EFL context.

A Rasch PCA of item residuals analysis was conducted to investigate the dimensionality of the Desire to Speak English measure. The results showed that 58.80% of the variance (eigenvalue = 12.90) was explained by the Rasch model and this percentage exceeded the 50.00% criterion proposed by Linacre (2007). In addition, 19.90% of the variance (eigenvalue = 4.40) was explained by the items. Finally, 11.90% of the variance (eigenvalue = 2.60) was unexplained by the first residual contrast. This figure did not meet the < 10.00% criterion, but the more important criterion of an eigenvalue < 3.00 was met. The results of the Rasch PCA of item residuals analysis indicated that the Desire to Speak English measure was fundamentally unidimensional.

Table 17 shows the loadings of nine Desire to Speak English items. Three items had loadings greater than 0.60: item D02, *I want to make friends with international students*, item D03, *I want to speak English with people from different countries*, and item D01, *I want to make friends with people from English-speaking countries*. These three items are associated with the participants' desire to socially reach out to people outside of Japan. Therefore, they are related to the social interaction facet. On the other hand, three items had negative loadings greater than -0.40: item D08, *I want to speak English as much as possible in English class*, item D09, *I want to find opportunities to speak English outside*

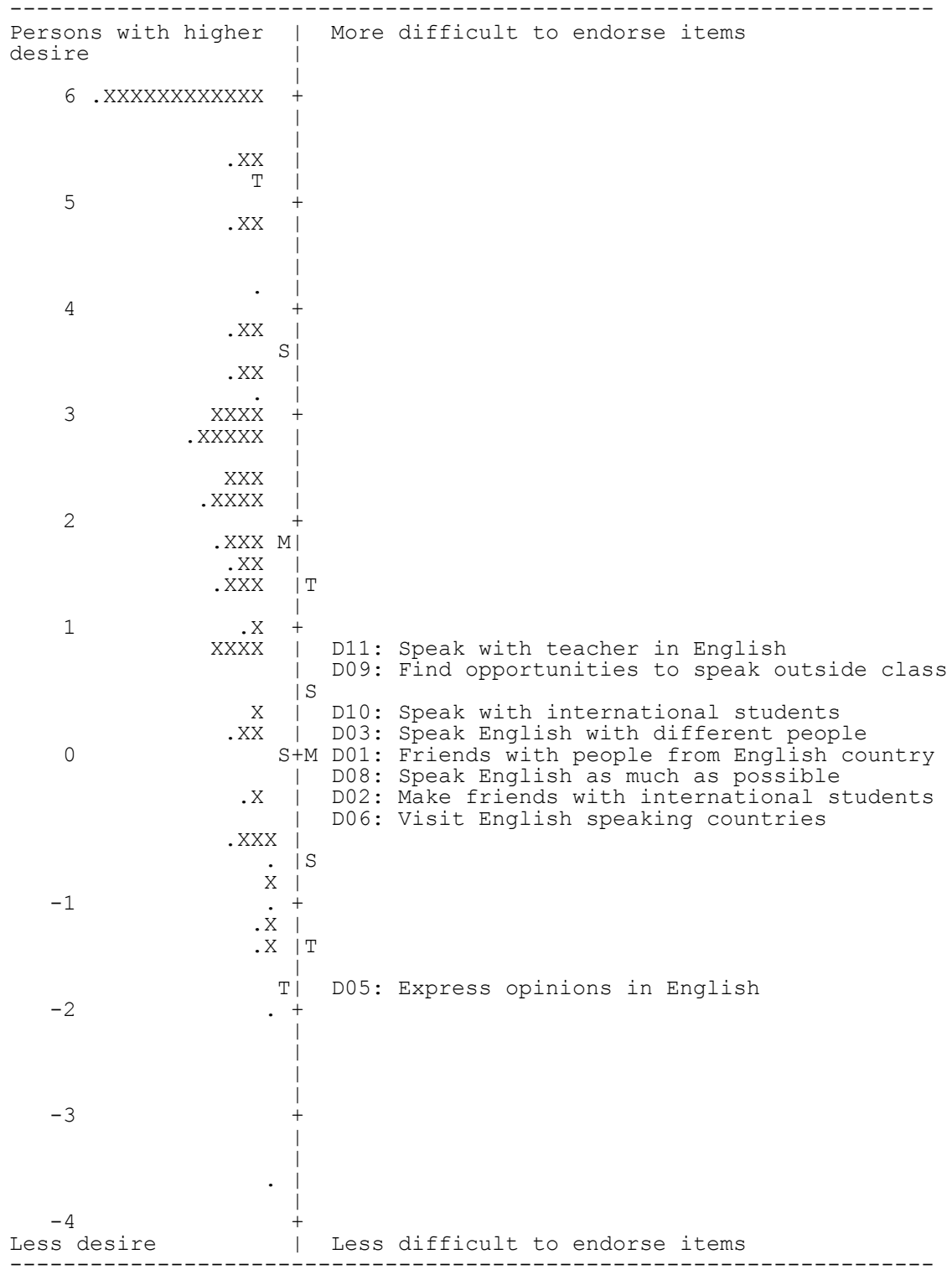



*Figure 6*. Wright map for desire to speak English. M = Mean; S = 1 SD; T = 2SDs. Each X = 2 participants.

*class*, and item D11, *I want to speak with the English teacher in English*. These three items concern the participants' desire to seek opportunities to practice English to better their current selves. In sum, it can be considered that the Desire to Speak English construct was fundamentally unidimensional with two facets. The first facet is related to socially communicate with people and the second facet is related to ambition to improve and to excel at speaking English.

Table 17. *Rasch PCA of Item Residuals for Desire to Speak English*

| Item | Loading | Measure | Infit MNSQ | Outfit MNSQ |
|---|---|---|---|---|
| D02 Make international student friends | 0.80 | -0.14 | 0.77 | 0.71 |
| D03 Speak with different country people | 0.69 | 0.17 | 0.75 | 0.73 |
| D01 Make other country friends | 0.68 | -0.03 | 1.03 | 1.15 |
| D10 Speak with international students | 0.14 | 0.33 | 0.61 | 0.63 |
| D08 Speak much English as possible | -0.68 | 0.08 | 1.03 | 1.09 |
| D09 Seek opportunities out of class | -0.56 | 0.71 | 0.95 | 0.91 |
| D11 Speak with teacher in English | -0.40 | 0.78 | 1.00 | 1.04 |
| D06 Visit English-speaking countries | -0.18 | -0.17 | 1.42 | 1.51 |
| D05 Express opinions and ideas | -0.17 | -1.72 | 1.39 | 1.66 |

The Rasch item reliability (item separation) was .96 (5.08), and the Rasch person reliability (separation) was .85 (2.39); indicating that the items were reliable. The participants endorsed the items strongly and displayed moderately high reliability. According to the separation statistics, the items formed over five distinct levels of difficulty, while the persons formed approximately two distinct levels.

**Descriptive Statistics for Desire to Speak English**

After conducting the Rasch analysis of the Desire to Speak English measure, the Rasch person ability estimates were exported into SPSS in order to check assumption of

normality. There were no outliers in this measure with a *z*-score above ±3.29 (Field, 2009). The skewness of 0.31 indicated that the data were slightly skewed and the kurtosis value of -0.71 indicated that the distribution was flatter than a normal distribution with shorter tails. Table 18 shows that both skewness and kurtosis values were under ±2.00 so the measure was assumed to be normally distributed (George & Mallery, 2016).

Table 18. *Descriptive Statistics for Desire to Speak English Items*

| | |
|---|---|
| *M* | 2.60 |
| *SE* | 0.17 |
| 95% CI | [2.25, 2.95] |
| *SD* | 2.51 |
| Skewness | 0.31 |
| *SES* | 0.17 |
| Kurtosis | -0.74 |
| *SEK* | 0.34 |

*Note*. *N* = 208; SES = Standard error skewness; SEK = Standard error kurtosis.

Three Rasch analyses were conducted to validate the three constructs: Classroom English Speaking Anxiety, English Speaking Self-efficacy, and Desire to Speak English constructs. Overall, the outcomes were satisfactory. All three constructs were fundamentally unidimensional, they displayed good reliability and separation indices, and the item fit statistics were acceptable.

### Preliminary Analyses of Self-Assessment of Speaking Task Questionnaire

The Self-Assessment of Speaking Task Questionnaire was administered after the four-picture cartoon narration pretest and posttest to measure changes in the learners' self-assessment of their own narration task performance in terms of syntactic complexity, syntactic accuracy, and oral fluency. Rasch analyses were conducted for the three latent

constructs to examine item fit to the Rasch model, construct dimensionality, reliability, and in order to produce interval measures. Data collected from the pretest and posttest questionnaires were stacked to ensure that the same item difficulty estimates and rating scale were used at both times; thus, any differences were due to changes in the participants' responses. Separate preliminary analyses of the pretest and posttest data showed that the questionnaire items displayed acceptable fit to the Rasch model.

**Rasch Analysis of Item Fit and PCA of Item Residuals for Self-Assessed Syntactic Complexity**

Similar to the above analyses, WINSTEPS was also used to analyze the raw data collected from the Self-Assessment of Speaking Task Questionnaire. Five items measuring the participants' self-assessment of syntactic complexity were analyzed. Data collected from questionnaires administrated after the pretest and posttest narrations were stacked into a single data set. Each person was considered a unique individual at each administration; therefore, data from 208 participants were analyzed (i.e., 104 participants x 2 administrations = 208). Responses on the six-point Likert-scale (1 = *Strongly disagree*, 2 = *Disagree*, 3 = *Slightly disagree*, 4 = *Slightly agree*, 5 = *Agree*, and 6 = *Strongly agree*) were examined (Table 19). In the original category function analysis, the outfit MNSQ statistic of Category 6 was slightly over the 2.00 criterion at 2.01. However, it had only one observation, so it did not meet the minimum of 10 observations per category; therefore, Category 6 was combined with Category 5 and the analysis was rerun. The new Category 5 (*Agree*) had 16 observations (2.00%) with an outfit MNSQ

statistic of 1.34, which was below the 2.00 criterion (Linacre, 2002b). The outfit MNSQ statistics of the remaining five categories were below 2.00 and the separation between the five categories was greater than the minimum threshold of 0.81 (Wolfe & Smith, 2007). The smallest separation of 1.77 was found between Category 1 (-4.46) and Category 2 (-2.23). Therefore, it was concluded that the performance of the new five-point Likert-scale was satisfactory and all categories functioned well.

Table 19. *Category Structure Functioning for Self-Assessed Syntactic Complexity*

| Category | Count | (%) | Infit MNSQ | Outfit MNSQ | Structure measure | Category measure |
|---|---|---|---|---|---|---|
| 1 Strongly disagree | 233 | 22 | 0.95 | 0.95 | None | (-4.46) |
| 2 Disagree | 337 | 32 | 0.93 | 0.92 | -3.28 | -2.23 |
| 3 Slightly disagree | 322 | 31 | 0.84 | 0.83 | -1.09 | -0.02 |
| 4 Slightly agree | 131 | 13 | 1.21 | 1.23 | 1.05 | 2.22 |
| 5 Combine Agree | 16 | 1 | 1.36 | 1.34 | 3.33 | (4.50) |

Table 20 displays the Rasch item statistics for the self-assessed syntactic complexity measure. All items met the infit MNSQ criterion of 0.50 to 1.50 (Linacre, 2002a). Item C4, *I could speak using clauses*, had the highest infit MNSQ of 1.32 and item C3, *I could speak using a combination of various sentences*, had the lowest infit MNSQ of 0.78.

Table 20. *Rasch Item Statistics for Self-Assessed Syntactic Complexity*

| Item | Measure | *SE* | Infit MNSQ | Infit ZSTD | Outfit MNSQ | Outfit ZSTD | Pt-measure correlation |
|---|---|---|---|---|---|---|---|
| C1: Complex grammar | 0.90 | 0.12 | 0.81 | -2.00 | 0.78 | -2.30 | .77 |
| C5: Linguistic features | 0.16 | 0.11 | 1.25 | 2.30 | 1.25 | 2.40 | .70 |
| C2: Long sentences | -0.01 | 0.11 | 0.82 | -1.90 | 0.81 | -12.00 | .82 |
| C3: Variety of sentences | -0.15 | 0.11 | 0.78 | -2.30 | 0.80 | -2.10 | .79 |
| C4: Clauses | -0.90 | 0.11 | 1.32 | 2.90 | 1.31 | 2.90 | .74 |
| *M* | 0.00 | 0.11 | 0.99 | -0.20 | 0.99 | -0.20 | — |
| *SD* | 0.58 | 0.00 | 0.24 | 2.30 | 0.24 | 2.40 | — |

The Wright map (Figure 7) presents the participants on the left according to their level of self-assessed syntactic complexity of the narrative task and the items on the right according to their difficulty. The left side of Figure 7 shows that most participants were well spread out between 0 to -4 logits, which indicates that they had negative perceptions of their own speaking complexity. This finding supports the generally humble character of the Japanese participants, but it could also indicate that the participants do not use complex sentences. A few participants were located at + 1 and + 2 logits, indicating that they believed that they could produce moderately complex utterances. On the other hand, there is a large cluster of participants located at -5 logits, which indicates that they had an extremely low perception of their ability to produce complex sentences. The right side of Figure 7 shows that item C1, *I could speak using difficult grammar*, is the most difficult item to endorse. The next most difficult to endorse item was item C5, *I could speak using unfamiliar sentences*, and it was followed by item C2, *I could speak using long sentences.* Both using new linguistic features and speaking in long sentences are difficult for most Japanese students where speaking English is concerned. Learners have the tendency to use known structures and to use relatively short sentences when speaking because speaking is a cognitively demanding on-line task. The easiest to endorse item was item C2, *I could speak using clauses*. This item was probably easy to endorse because the target students have knowledge of basic clauses and can produce sentences such as *I think that speaking is difficult*. Overall, the five items had similar difficulty estimates and they clustered around the mean.

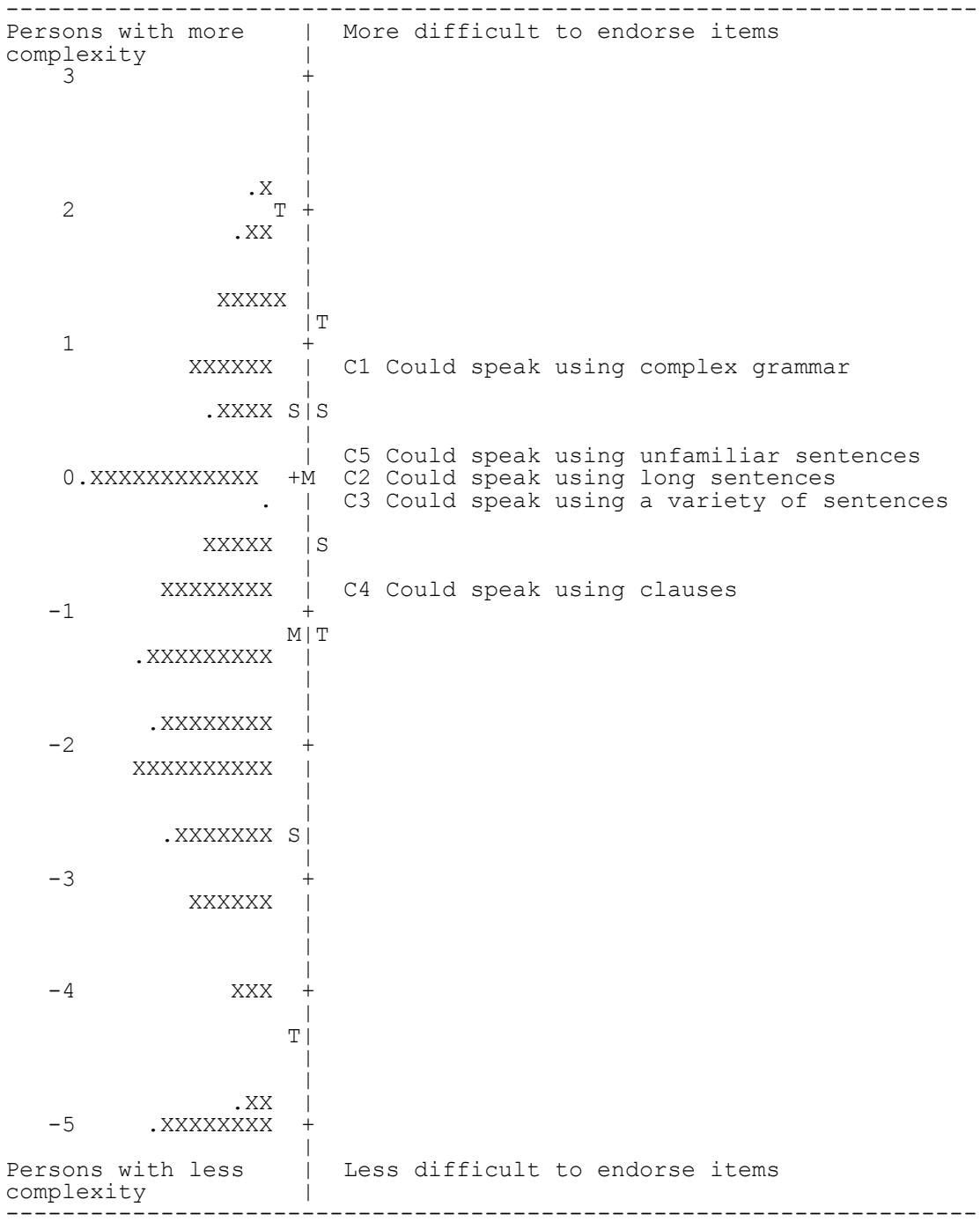



*Figure 7*. Wright map for self-assessed syntactic complexity. M = Mean; S = 1 SD; T = 2SDs. Each X = 2 participants. Each . = 1 participant.

A Rasch PCA of item residuals analysis was conducted to investigate the dimensionality of the self-assessed syntactic complexity measure. The results showed

that 54.90% of the variance (eigenvalue = 6.10) was explained by the Rasch model. This percentage exceeded the 50.00% criterion suggested by Linacre (2007). In addition, 19.60% of the variance (eigenvalue = 2.20) was explained by the items. Finally, the first residual contrast had 14.00% of unexplained variance. Although this result did not meet the < 10.00% criterion, the eigenvalue of 1.50 met the < 3.00 criterion. Overall, the PCA results indicated that the five items were fundamentally unidimensional.

Table 21 shows the loading of the five self-assessed syntactic complexity items. Items C2 (0.71) and C1 (0.67) had high positive loadings; both items concerned the general use of grammar. In contrast, items C4 (-0.66) and C5 (-0.41) had high negative loadings. However, it is difficult to categorize Items C4 and C5 into a distinct facet despite their high negative loadings.

Table 21. *Rasch PCA of Item Residuals for Self-Assessed Syntactic Complexity*

| Item | Loading | Measure | Infit MNSQ | Outfit MNSQ |
|---|---|---|---|---|
| C2: Long sentences | 0.71 | -0.01 | 0.82 | 0.81 |
| C1: Complex grammar | 0.67 | 0.90 | 0.81 | 0.78 |
| C3: Variety of sentences | -0.04 | -0.15 | 0.78 | 0.80 |
| C4: Clauses | -0.66 | -0.90 | 1.32 | 1.31 |
| C5: Unfamiliar sentences | -0.41 | 0.16 | 1.25 | 1.25 |

Rasch item reliability (item separation) was .96 (4.81), and the Rasch person reliability (separation) was .79 (1.94), indicating that the items were highly reliable. On the other hand, the Rasch person reliability coefficient just missed the .80 benchmark and the separation index just missed the 2.00 criterion; thus, the participants' ability estimates were moderately reliable. The items formed more than four distinct levels of difficulty, while the persons formed approximately two distinct levels.

**Descriptive Statistics for Self-Assessed Syntactic Complexity**

After conducting the Rasch analysis of the self-assessed syntactic complexity measure, the person ability estimates from the Rasch analysis were exported into SPSS in order to confirm the normal distribution of the person measures (Table 22). There were no outliers in this measure with *z*-score above ±3.29 (Field, 2009). The skewness statistic of -0.64 indicated that it was slightly skewed and kurtosis was -0.09, which indicated a normal distribution, as they were both under ±2.00 (George & Mallery, 2016).

Table 22. *Descriptive Statistics for Self-Assessed Syntactic Complexity Items*

| | |
|---|---|
| *M* | -1.56 |
| *SE* | 0.14 |
| 95% CI | [-1.84, -1.28] |
| *SD* | 2.05 |
| Skewness | -0.64 |
| *SES* | 0.17 |
| Kurtosis | -0.09 |
| *SEK* | 0.34 |

*Note*. *N* = 208; SES = Standard error skewness; SEK = Standard error kurtosis.

**Rasch Analysis of Item Fit and PCA of Item Residuals for Self-Assessed Syntactic Accuracy**

The items measuring the learners' perceptions of their own English speaking accuracy were analyzed by conducting a Rasch analysis with the responses collected from the Self-Assessment of Speaking Task Questionnaire administered after the pretest and posttest narrations. The data were stacked into one data set; thus, data from 208 participants were analyzed (i.e., 104 participants x 2 administrations = 208). Responses were first examined by checking category functioning (Table 23). The minimum of 10

observations per category for a six-point scale was met (Linacre, 2002a). The smallest number of observations was for Category 6 (*Strongly agree*), which had 33 observations (2.00%) and the largest number of observations was for Category 4 (*Slightly agree*), which had 570 observations (39.00%). The outfit MNSQ statistics for all six categories were below the 2.00 criterion (Linacre, 2002b). The separation between the categories was greater than the required 0.59 logits for a six-point scale (Linacre, 2002b); thus, it was concluded that the six-point Likert-scale performed satisfactorily.

Table 23. *Category Structure Functioning for Self-Assessed Syntactic Accuracy*

| Category | Count | (%) | Infit MNSQ | Outfit MNSQ | Structure measure | Category measure |
|---|---|---|---|---|---|---|
| 1 Strongly disagree | 132 | 9 | 1.02 | 1.06 | None | (-4.23) |
| 2 Disagree | 207 | 14 | 1.03 | 1.06 | -2.91 | -2.58 |
| 3 Slightly disagree | 365 | 25 | 0.96 | 0.96 | 1.94 | -1.26 |
| 4 Slightly agree | 570 | 39 | 0.89 | 0.91 | -0.83 | 0.65 |
| 5 Agree | 143 | 10 | 1.00 | 0.96 | 2.09 | 2.87 |
| 6 Strongly agree | 33 | 2 | 1.27 | 1.24 | 3.59 | (4.83) |

Item A7 had an infit MNSQ statistic of 1.56, which exceeded the .50-1.50 benchmark. In addition, item A7, *I could speak using accurate verbs and nouns*, concerned the participants' ability to distinguish the use of verbs and nouns, which concerns lexical accuracy rather than syntactic accuracy. Therefore, a decision was made to remove item A7 from the analysis. Table 24 shows the item statistics for the Self-Assessed Syntactic Accuracy items after deleting item A7. All of the remaining items met the 0.50-1.50 criterion suggested by Linacre (2002a). In order to check the effect of deleting item A7, the Rasch person measures with and without item A7 were exported

into SPSS and a Pearson correlation analysis was run. The Pearson correlation of .99 ($p < .01$) indicated that deleting the item did not substantially affect person measures.

Table 24. *Rasch Item Statistics for Self-Assessed Syntactic Accuracy*

| Item | Measure | *SE* | Infit MNSQ | Infit ZSTD | Outfit MNSQ | Outfit ZSTD | Pt-measure correlation |
|---|---|---|---|---|---|---|---|
| A5: Past perfect | 1.54 | 0.10 | 1.00 | 0.00 | 1.01 | 0.10 | .71 |
| A4: Past continuous | 0.64 | 0.10 | 1.03 | -0.30 | 1.04 | 0.40 | .77 |
| A6: Articles | -0.18 | 0.10 | 1.49 | 4.20 | 1.48 | 4.00 | .70 |
| A1: Grammar | -0.23 | 0.10 | 0.85 | -1.50 | 0.86 | -1.40 | .73 |
| A2: Verb tenses | -0.50 | 0.10 | 0.72 | -2.90 | 0.73 | -2.80 | .81 |
| A3: Simple past | -1.26 | 0.11 | 0.92 | -0.70 | 0.85 | 1.40 | .78 |
| *M* | 0.00 | 0.10 | 1.00 | -0.10 | 1.00 | -0.20 | — |
| *SD* | 0.88 | 0.00 | 0.24 | 2.20 | 0.24 | 2.10 | — |

The Wright map (Figure 8) presents the participants on the left according to their level of self-assessed accuracy of the narrative task and the items on the right according to their difficulty estimates. Figure 8 shows that most participants clustered around the mean at +1.0 to -2.0 logits. This result indicates that the participants generally had a slightly low perception of their speaking accuracy. There were a few outliers located around +5.0 logits suggesting that they had very high estimates of their ability to speak accurately. On the other hand, a small group of participants located at -6.0 logits indicated that had very low perceptions of their speaking accuracy. The right side of Figure 8 shows that the most difficult to endorse item was item A5, *I could speak using accurate past perfect tense*. Most Japanese have difficulty acquiring past perfect tense and rarely, if ever, use it when speaking. The participants' perceptions of their ability to produce past perfect tense accurately is in accord with their oral production, as past perfect was used a very small number of times in the participant narrations.

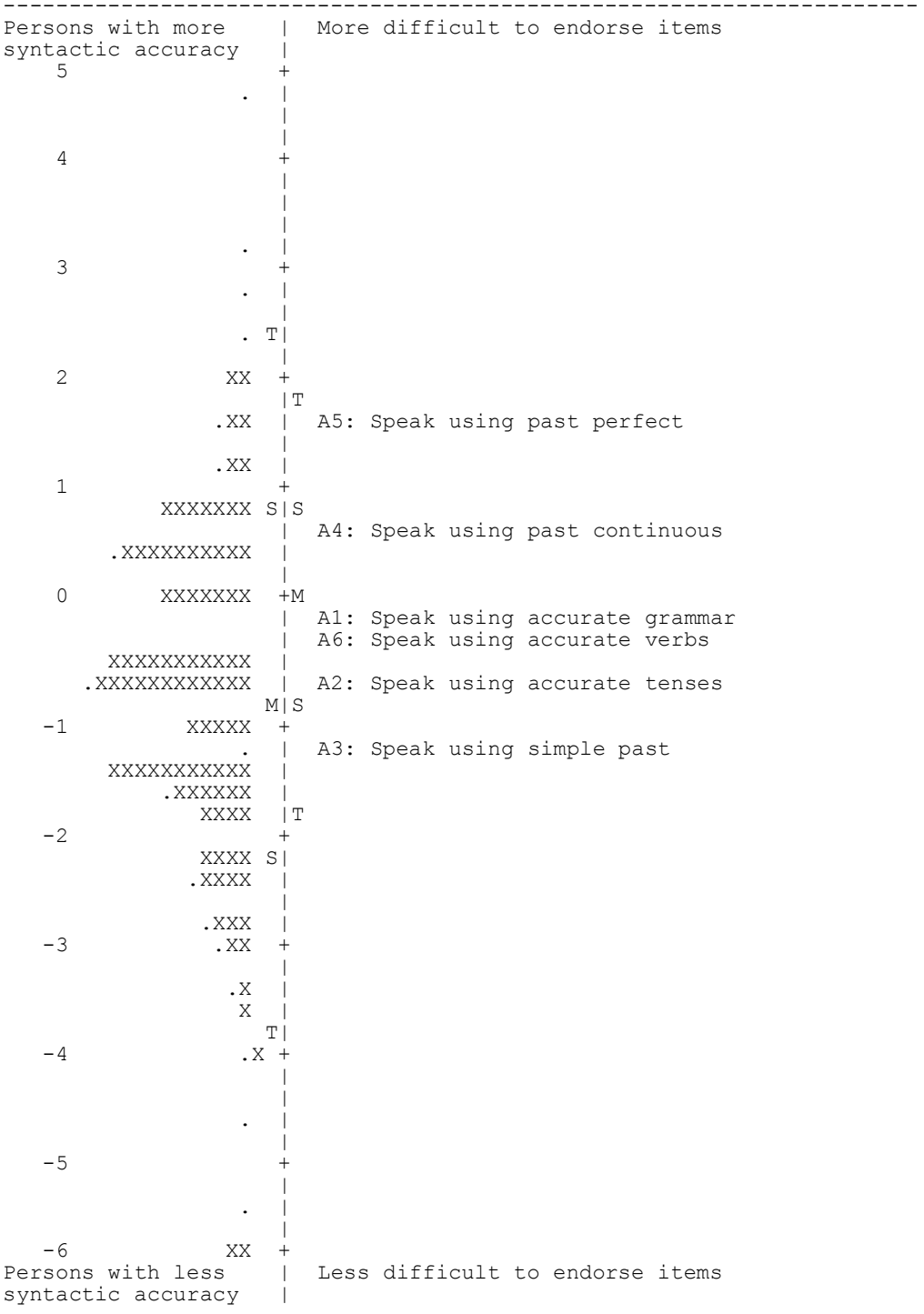



*Figure 8*. Wright map for self-assessed syntactic accuracy. P = person. M = Mean; S = 1 SD; T = 2SDs. Each X = 2 participants.

The second most difficult to endorse item was item A4, *I could speak using accurate past continuous tense*. The easiest item to endorse was item A3, *I could speak using accurate simple past tense*. Most Japanese students learn simple past tense in junior high school and by the time they are in university, they are generally able to use it accurately.

The Rasch item reliability estimate (item separation) was .99 (8.43), and the Rasch person reliability estimate (separation) was .82 (2.17); indicating that the items were highly reliable. On the other hand, the participants' ability estimates were moderately reliable. In addition, the items formed over eight distinct levels of difficulty, while the persons formed approximately two distinct levels according to the Rasch separation statistics.

A Rasch PCA of item residuals analysis was conducted to investigate the dimensionality of the self-assessed syntactic accuracy measure. The results showed that 61.10% of the variance (eigenvalue = 9.30) was explained by the Rasch model; this percentage exceeded the 50.00% criterion by Linacre (2007). In addition, 35.80% of the variance (eigenvalue = 5.50) was explained by the items. Finally, 12.10% of the variance (eigenvalue = 1.80) was unexplained by the first residual contract. Although it was not below the 10.00% benchmark, it met the < 3.00 eigenvalue criterion. The results of the Rasch PCA of item residuals analysis indicate that the self-assessed accuracy measure is fundamentally unidimensional.

Table 25 shows the loadings of the self-assessed syntactic accuracy items. Four items had loadings greater than 0.40 or -0.40. Item A5, *I could speak using accurate past perfect tense*, and item A4, *I could speak using accurate past continuous tense*, had high

positive loadings. Item A2, *I could speak using accurate verb tenses*, and item A1, *I could speak using accurate grammar*, had high negative loadings. Items A5 and A4 are related to specific grammatical features, whereas items A1 and A2 concern general grammatical ability. Overall, the construct was fundamentally unidimensional.

Table 25. *Rasch PCA of Item Residuals for Self-Assessed Syntactic Accuracy*

| Item | Loading | Measure | Infit MNSQ | Outfit MNSQ |
|---|---|---|---|---|
| A5: Past perfect tense | 0.74 | 1.54 | 1.00 | 1.01 |
| A4: Past continuous tense | 0.73 | 0.64 | 1.03 | 1.04 |
| A2: Accurate verb tenses | -0.63 | -0.50 | 0.72 | 0.73 |
| A1: Accurate grammar | -0.47 | -0.23 | 0.85 | 0.86 |
| A6: Accurate articles | -0.32 | -0.18 | 1.49 | 1.48 |
| A3: Simple past tense | -0.24 | -1.26 | 0.92 | 0.85 |

**Descriptive Statistics for Self-Assessed Syntactic Accuracy**

After conducting the Rasch analysis of the self-assessed syntactic accuracy measure, the Rasch person ability estimates were exported into SPSS in order to confirm the normal distribution of responses (Table 26). There were no outliers in this measure with a *z*-score above ±3.29 (Field, 2009). The descriptive statistics in Table 14 show that skewness was -0.38 and kurtosis was 1.27. Both skewness and kurtosis were under ±2.00, which indicated a reasonably normal distribution (George & Mallery, 2016).

Table 26. *Descriptive Statistics for Self-Assessed Syntactic Accuracy Items*

| | |
|---|---|
| *M* | -0.85 |
| *SE* | 0.12 |
| 95% CI | [-1.08, -0.61] |
| *SD* | 1.72 |
| Skewness | -0.38 |
| *SES* | 0.17 |
| Kurtosis | 1.27 |
| *SEK* | 0.34 |

*Note*. *N* = 208; SES = Standard error skewness; SEK = Standard error kurtosis.

**Rasch Analysis of Item Fit and PCA of Item Residuals for Self-Assessed Oral Fluency**

A total of seven items measuring the participants' self-assessment of fluency performance for the narrative task were administrated after the pretest and posttest narrations. Data from both administrations were stacked into a single data set in which each person was considered a unique individual at each administration. Therefore, data from 208 participants were analyzed (i.e., 104 participants x 2 administrations = 208).

The performance of the six-point Likert-scale was first examined (Table 27). In the original category function analysis, Category 6 (*Strongly agree*) had only six observations, which did not meet the minimum 10 observations per category criterion (Linacre, 2002b). Therefore, it was combined with Category 5 and the analysis was rerun. The new Category 5 (*Combine agree*) had 43 observations (3.00%) with an outfit MNSQ statistic of 1.41. The separation between the five categories was greater than the minimum threshold of 0.81 logits (Wolfe & Smith, 2007). In addition, the outfit MNSQ statistics of the remaining five categories were all below the 2.00 criterion. Therefore, the new five-category scale was considered to be appropriate and used for further analyses.

Table 27. *Category Structure Functioning for Self-Assessed Oral Fluency*

| Category | Count | (%) | Infit MNSQ | Outfit MNSQ | Structure Measure | Category measure |
|---|---|---|---|---|---|---|
| 1 Strongly disagree | 252 | 17 | 0.88 | 0.89 | None | (-3.94) |
| 2 Disagree | 351 | 24 | 0.98 | 0.97 | -2.68 | -2.06 |
| 3 Slightly disagree | 525 | 36 | 0.83 | 0.81 | -1.32 | -0.22 |
| 4 Slightly agree | 282 | 19 | 1.07 | 1.06 | 0.80 | 2.04 |
| 5 Combine agree | 43 | 3 | 1.55 | 1.41 | 3.21 | (4.37) |

Table 28 shows the Rasch item statistics for the self-assessed oral fluency construct after deleting items F4 and F6. All of the items except F4, *I could speak without repetitions*, and F6, *I could speak without reconstructing sentences*, met the 0.50-1.50 criterion suggested by Linacre (2002a). Item F4 had an infit MNSQ statistic of 1.83, an outfit MNSQ statistic of 1.83. Item F6 had an infit MNSQ statistic of 1.62, an outfit MNSQ statistic of 1.62. In order to justify the exclusion of items F4 and F6, the Rasch person measures with and without items F4 and F6 were exported into SPSS and a bivariate correlation analysis was conducted. The Pearson correlation of .95 ($p < .01$) indicated that deleting the two items did not substantially affect person measures. Therefore, F4 and F6 were deleted from further analyses.

Table 28. *Rasch Item Statistics for Self-Assessed Oral Fluency*

| Item | Measure | *SE* | Infit MNSQ | Infit ZSTD | Outfit MNSQ | Outfit ZSTD | Pt-measure correlation |
|---|---|---|---|---|---|---|---|
| F3 | 0.29 | 0.13 | 0.82 | -1.80 | 0.83 | -1.60 | .85 |
| F2 | 0.22 | 0.13 | 0.88 | -1.20 | 0.90 | -0.90 | .86 |
| F5 | 0.06 | 0.13 | 1.48 | 4.10 | 1.44 | 3.80 | .80 |
| F1 | -0.08 | 0.13 | 0.51 | -5.90 | 0.49 | -6.10 | .92 |
| F7 | -0.48 | 0.13 | 1.27 | 2.50 | 1.23 | 2.20 | .81 |
| *M* | 0.00 | 0.13 | 0.99 | -0.50 | 0.98 | -0.50 | — |
| *SD* | 0.27 | 0.00 | 0.34 | 3.50 | 0.33 | 3.40 | — |

The Wright map (Figure 9) presents the participants on the left according to their level of self-assessed oral fluency performance of the narrative task and the items on the right according to their endorsability. The left side of Figure 9 shows that most participants were well spread out, but the majority were below 0 logits, indicating that they had a slightly low perception of their ability to speak English fluently. There were a few clusters near -5.00 and -6.00, indicating that some participants have a very low

perception of their speaking fluency. The right side of Figure 9 shows all items clustered around the mean indicating that they were all moderately easy to endorse by the target participants. The items formed three clusters; the first cluster has items F2, F3, F1 and F5, item F7 was by itself. Item F2, *I could speak with few pauses (without having much interval)*, was the most difficult item to endorse, as pauses are a part of natural human speech and it is difficult for participants to speak without pausing. The next most difficult items were item F3, *I could speak speedily*, item F1, *I could speak smoothly*, and item F5, *I could speak without stuttering*. These three items also concern the nature of human speech, as it is naturally difficult for some participants to speak rapidly, smoothly, and without stuttering even in their first language. The easiest item to endorse was item F4, *I could speak without repetitions*.

A Rasch PCA of item residuals analysis was conducted to investigate the dimensionality of the self-assessed fluency measure. The Rasch model explained 65.90% of the variance (eigenvalue = 9.60); this percentage exceeded the 50.00% criterion (Linacre, 2007). In addition, 13.50% of the variance (eigenvalue = 2.00) was explained by the items. Finally, 12.70% of the variance (eigenvalue = 1.90) was unexplained by the first residual contract. Although it did not meet the 10.00% criterion, it met the < 3.00 eigenvalue criterion. Therefore, the results of the Rasch PCA of item residuals analysis indicated that the self-assessed fluency measure was fundamentally unidimensional.

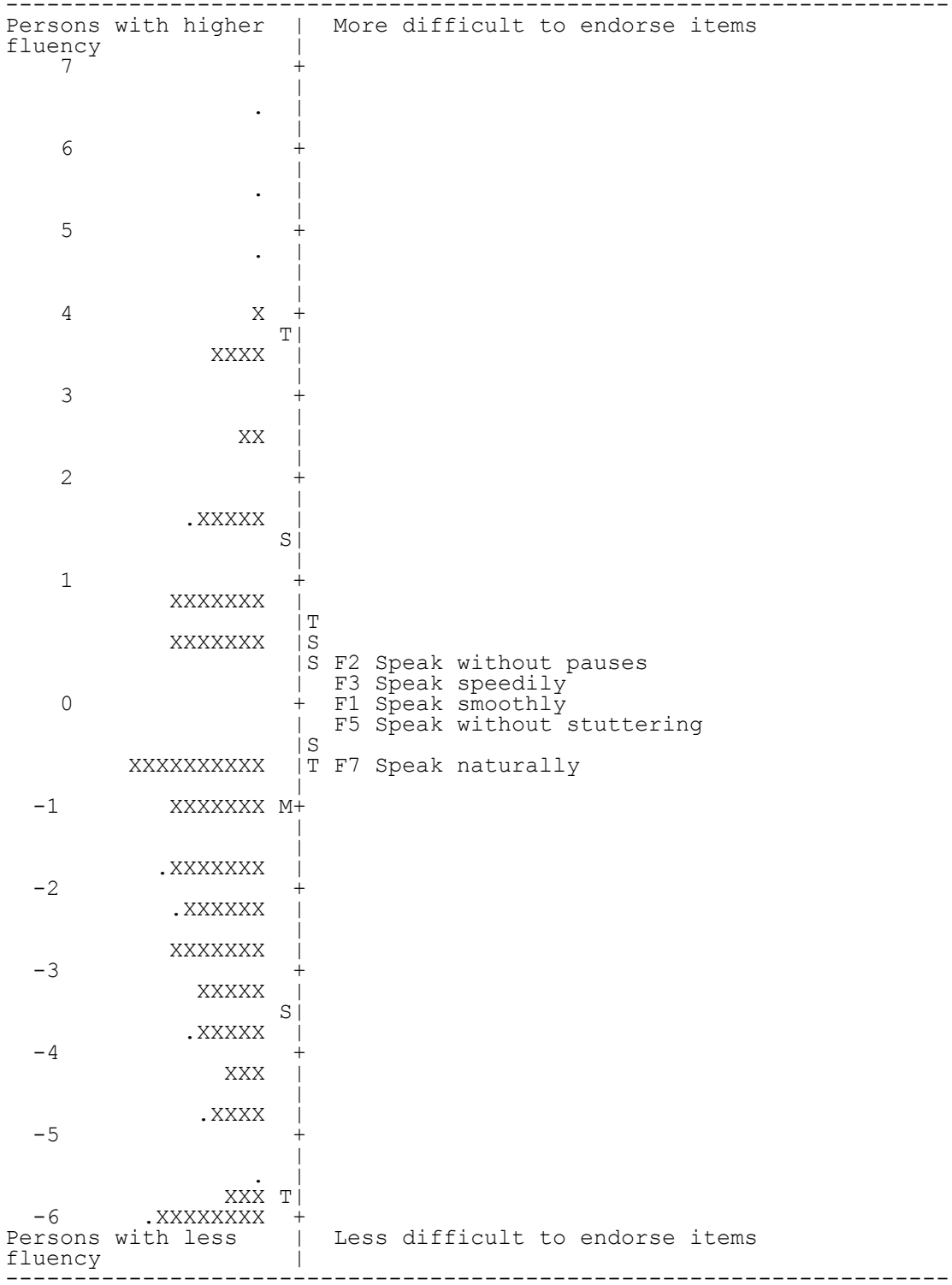



*Figure 9*. Wright map for self-assessed oral fluency. M = Mean; S = 1 SD; T = 2SDs. Each X = 2 participants. Each . = 1 participant.

Table 29 shows the loadings of the five self-assessed fluency items. Three items had loadings greater than 0.40: Item F1, *I could speak smoothly*, item F2, *I could speak with few pauses (without having much interval*, and item F3, *I could speak speedily*. These three ideas are related to speech rate. On the other hand, items F5, *I could speak without stuttering*, and item F7, *I could speak naturally*, had loadings greater than -0.40. Items F5 and F7 are more related to speaking style. Overall, the results indicated that the self-assessed oral fluency construct is fundamentally unidimensional.

Table 29. *Rasch PCA of Item Residuals for Self-Assessed Oral Fluency*

| Item | Loading | Measure | Infit MNSQ | Outfit MNSQ |
|---|---|---|---|---|
| F1 Speak smoothly | 0.50 | -0.08 | 0.51 | 0.49 |
| F2 Speak without pauses | 0.71 | 0.22 | 0.88 | 0.90 |
| F3 Speak speedily | 0.52 | 0.29 | 0.82 | 0.83 |
| F7 Speak naturally | -0.69 | 0.06 | 1.48 | 1.44 |
| F5 Speak without stuttering | -0.60 | -0.48 | 1.27 | 1.23 |

The Rasch item reliability estimate (item separation) was .78 (1.77); the Rasch person reliability estimate (separation) was .87 (2.54). Item reliability was slightly low, as it was below the .80 benchmark. The items' similar difficulty estimates resulted in the moderate reliability. The person reliability was good. The separation statistics indicated that the items and persons formed approximately two statistically distinct levels.

**Descriptive Statistics for Self-Assessed Oral Fluency**

After the Rasch analysis of the self-assessed fluency measure, the person ability estimates from the Rasch analysis were exported into SPSS in order to confirm the normal distribution of responses (Table 30). There were no outliers in this measure with

*z*-score above ±3.29 (Field, 2009). The skewness statistic of -0.09 and the kurtosis statistic of -0.12 met ±2.00 criterion indicating a normal distribution (Field, 2009; (George & Mallery, 2016).

Table 30. *Descriptive Statistics for Self-Assessed Oral Fluency Items*

| | |
|---|---|
| *M* | -1.56 |
| *SE* | 0.20 |
| 95% CI | [-1.95, -1.18] |
| *SD* | 2.81 |
| Skewness | -0.09 |
| *SES* | 0.17 |
| Kurtosis | -0.12 |
| *SEK* | 0.34 |

*Note*. *N* = 208; SES = Standard error skewness; SEK = Standard error kurtosis

Overall, the self-assessed syntactic complexity, syntactic accuracy, and oral fluency constructs on the Self-assessment of Speaking Task Questionnaire achieved satisfactory outcomes. The Likert scales performed well, item fit was satisfactory, the constructs were fundamentally unidimensional, and the item and person reliability and separation statistics were satisfactory. The self-assessed oral fluency construct had slightly low item reliability and non-normal distributions, but was still considered acceptable. The weaker item reliability of the self-assessed oral fluency construct was a limitation of the study; caution will be paid when interpreting the results of that construct.

## Multifaceted Rasch Analysis

Three human raters rated the pretest and posttest performances of the four-picture cartoon task of 104 candidates by analytically rating their oral proficiency on the pretest and posttest in terms of complexity, accuracy, and fluency. Multifaceted Rasch analysis

was conducted to examine rater severity, the performance of the rating scale, and participant performance in order to produce fair measure averages to reliably measure the participants' speaking proficiency. The following five-point Likert-scale was used: 1 = *Very low proficiency*; 2 = *Low proficiency*; 3 = *Neutral proficiency*; 4 = *High proficiency*; 5 = *Very high proficiency*. The rating data were analyzed using FACETS version 3.71.4 (Linacre, 2013). The initial analyses of the pretest and posttest data before data stacking showed that the items displayed acceptable fit to the Rasch model. The pretest and posttest data were then stacked into a single data set in which each person was considered a unique individual at each administration (Wright, 1996, 2003). Therefore, data of 208 participants were analyzed (i.e., 104 participants x two administrations = 208). Stacking ensured that the same rater severity, rating scale difficulty estimates, and task difficulty estimates were used at both rating occasions; thus, differences in the person ability estimates were due to changes in syntactic complexity, syntactic accuracy, and oral fluency, and not the result of changes in rater severity nor the rating.

The five-point rating scale was first examined using category functioning (Table 31). First, the minimum of 10 observations per category was met. Category 1 (*Very low proficient*) had the smallest number of observations, at 12 (1.00%). Second, the outfit MNSQ statistics for all categories were below 2.00. Separation between the categories was greater than the minimum requirement of 0.59 logits where the smallest separation was 2.03 logits between Category 3 (-1.07) and Category 4 (0.96). Based on the above results, it was concluded that the five-point rating system performed satisfactorily.

Table 31. *Category Structure Functioning for Analytical Ratings*

| Category | Count | (%) | Outfit MNSQ | Rash-Andrich thresholds | *SE* |
|---|---|---|---|---|---|
| 1 Very low proficient | 12 | 1 | 1.50 | None | — |
| 2 Low proficient | 133 | 12 | 0.90 | -3.74 | 0.32 |
| 3 Neutral proficient | 362 | 33 | 1.00 | -1.07 | 0.11 |
| 4 High proficient | 453 | 41 | 1.10 | 0.96 | 0.08 |
| 5 Very high proficient | 144 | 13 | 0.90 | 3.85 | 0.11 |

Figure 10 shows the map of the Rasch estimates for the rater ratings. The first column shows the logit scale ranging from -6 to +7, and the second column shows the persons with higher-proficiency participants at the top of the table. The third column shows rater severity with the strictest rater at the top, and the fourth column shows the difficulty of the rating criteria (syntactic complexity = Criterion 1, syntactic accuracy = Criterion 2, and oral fluency = Criterion 3), again with the most difficult criterion at the top of the figure. The final column shows the functioning of the five-point Likert scale.

Figure 10 indicates that most participants were around 0 to +2.0 logits, with a few outliers at the top of the table and only one outlier at the bottom. Rater 2 was the most severe rater followed by Rater 3 and Rater 1, with a very small difference in severity between Rater 3 and Rater 1. There was also a small range in the difficulty level of the three rating criteria. Oral fluency was slightly more difficult than syntactic accuracy, which was slightly more difficult than syntactic complexity. This result indicates that oral fluency was more difficult to endorse than syntactic accuracy and syntactic complexity for the raters, but the differences were small.

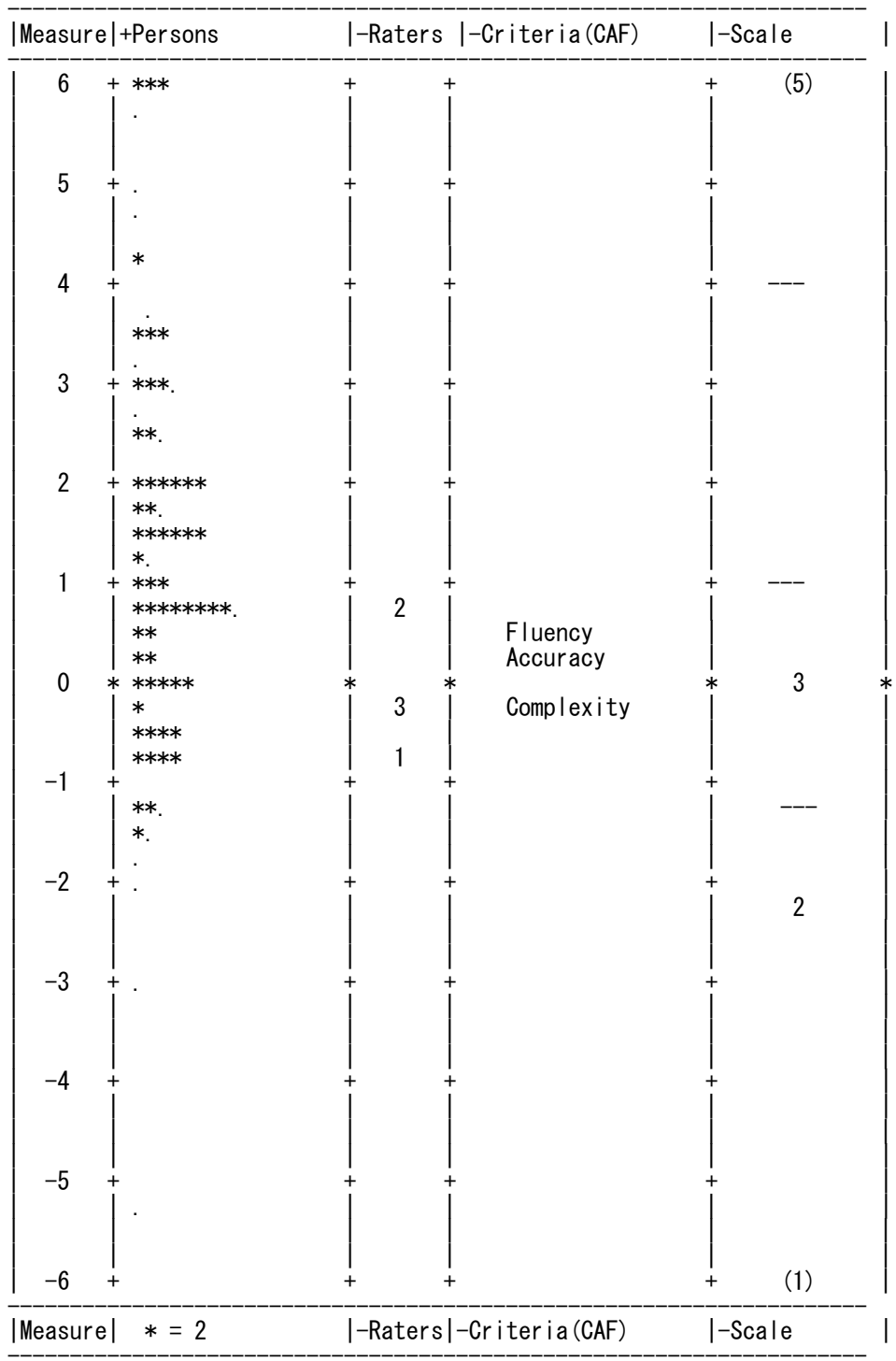



*Figure 10.* Facet map for stacked ratings. Each * = 2 persons.

Table 32 presents the Rasch fit statistics for the three criteria rated by the three raters. The model had a reliability estimate of .81, and item separation of 2.10. The person reliability estimate was .88, and the person separation statistic was 2.66. This model had acceptable separation indices and reliability values. All three criteria met the infit MNSQ criterion of 0.50 to 1.50 logits. According to Table 33, syntactic accuracy had the highest infit MNSQ of 1.18, but it was within the acceptable range. Oral fluency and syntactic complexity had similar infit MNSQ statistic of 0.93 and 0.89, respectively.

Table 32. *Criteria Difficulty Measurement Report for Analytical Ratings*

| Criterion | Observed count | Measure | *SE* | Infit MNSQ | Infit ZSTD |
|---|---|---|---|---|---|
| Accuracy | 379 | 0.25 | 0.08 | 1.18 | 2.2 |
| Fluency | 379 | 0.09 | 0.08 | 0.93 | -1.0 |
| Complexity | 379 | -0.33 | 0.09 | 0.89 | -1.5 |
| *M* | — | 0.00 | 0.08 | 1.00 | -0.1 |
| *SD* | — | 0.30 | 0.00 | 0.13 | 2.1 |

*Note*. Population separation = 2.71; Population reliability (reliable difference in the difficulty of items) = .88; Fixed (all same) chi-square = 24.8; *df* = 2, $p < .01$.

The results of the FACETS analysis for persons indicated that eight participants reached maximum infit MNSQ statistics. A closer examination of these participants indicated that they had all been given a maximum score (i.e., a rating of 15 by one rater or a rating of 30 by two raters). In addition, 16 persons had an infit MNSQ statistic of greater than 2.00 and one person had infit ZSTD statistic greater than 3.00. Although these persons did not satisfy Linacre's (2002a) benchmarks of infit MNSQ values < 2.00 and Infit ZSTD values < 3.00.

Table 33 presents the results of rater severity. Rater 2 was the strictest rater and Rater 1 was the most lenient rater. This result was similar to the pilot study where Rater 2

gave comparatively lower scores to pilot examinees compared to Raters 1 and 3. All three raters fit the Rasch model well using the 0.50 to 1.50 logits benchmark suggested by Linacre (2002a).

Table 33. *Rater Severity Measurement Report for Analytical Ratings*

| Rater | Observed count | Measure | Model *SE* | Infit MNSQ | Infit ZSTD |
|---|---|---|---|---|---|
| 2 | 300 | 0.87 | 0.09 | 0.95 | -0.60 |
| 3 | 303 | -0.13 | 0.09 | 0.90 | -1.20 |
| 1 | 534 | -0.75 | 0.07 | 1.11 | 1.60 |
| *M* | — | 0.00 | 0.09 | 0.99 | -0.10 |
| *SD* | — | 0.67 | 0.01 | 0.09 | 1.30 |

*Note.* Population Separation = 5.32; Population Reliability (reliable difference in severity of judges, not inter-rater) = .97; Fixed (all same) chi-square = 89.5, $df = 2$, $p < .01$.

After the Multifaceted Rasch analysis of rater analytical rating, logits were exported into SPSS in order to confirm the normal distribution of responses. Table 34 presents the descriptive statistics of the analytical ratings where pretest and posttest fair average logit statistics were calculated for each participant using multifaceted Rasch model. There were no outliers in this measure with a *z*-score above ±3.29 (Field, 2009). The descriptive statistics in Table 34 show that the normality of the distribution was assumed because the skewness and kurtosis statistics for the pretest and posttest for all three groups for were below ±2.0 (George & Mallery, 2016).

Overall, the three criteria of syntactic complexity, syntactic accuracy, and oral fluency as well as the three raters performed well. A decision was made to keep all examinees even though a number of examinees were given a maximum score by the raters and might not have been precisely measured by the model.

Table 34. *Descriptive Statistics of the Analytically Rated Pretest and Posttest*

| | Comparison group | | FFI group | | FFI + PF group | |
|---|---|---|---|---|---|---|
| | Pretest | posttest | Pretest | posttest | Pretest | posttest |
| *M* | 3.66 | 3.61 | 3.27 | 3.28 | 3.53 | 3.83 |
| *SE* | .14 | .12 | .11 | .09 | .13 | .08 |
| *95%CI* | [3.36, 3.95] | [3.37, 3.85] | [3.04, 3.49] | [3.10, 3.46] | [3.27, 3.78] | [3.65, 4.00] |
| *SD* | .72 | .59 | .71 | .57 | .78 | .52 |
| *SK* | .33 | .32 | .35 | .24 | -.35 | -.41 |
| *SES* | .46 | .46 | .37 | .37 | .38 | .38 |
| *KT* | -.70 | .30 | -.17 | -.09 | .96 | .29 |
| *SEK* | .90 | .90 | .72 | .72 | .75 | .75 |

*Note*. FFI = Form-focused instruction; FFI + PF = Form-focused instruction and peer feedback; CI = 95% confidence interval; SK = skewness; KT = kurtosis; SES = Standard error skewness; SEK = Standard error kurtosis. All statistics are based on Rasch logits.

## Three Types of Past Tense Forms Observed

The accuracy measurement in the present study was focused on the percentage of accurate and inaccurate use of simple past, past continuous, and past perfect verb tenses. The effectiveness of the intervention was measured by changes in the accurate and inaccurate usage ratio before and after the intervention. To make this measurement more valid, the frequency of the three past tense forms used was observed. In both the pretest and posttest, the participants predominantly used simple past form; very few past continuous forms, and almost no past perfect forms were used. The participants produced a total of 1,152 past tense forms on the pretest, 1,023 (89.00%) were simple past forms, 122 (11.00%) were past continuous forms, and 7 (< 0.05%) were past perfect forms. The participants produced a total of 1,304 past tense forms on the posttest, 1,115 (86.00%) were simple past forms, 171 (13.00%) were past continuous forms, and 18 (< 1.00%) were past perfect. Therefore, it was decided to exclude past perfect tense from the total count of past tense use from the present study. In addition, although past continuous forms made up approximately 10.00% of the total observations on both the pretest and

posttest, it was also excluded from the total count of past tense use because the four-picture cartoon prompt contains a textual cue that says, “one day a girl was watching a TV program,” which illustrates the use of past continuous. Therefore, out of the 122 instances of past continuous tense observed on the pretest, 102 of them used the textual cue given by the cartoon. In addition, out of the 171 instances of past continuous tense observed on the pretest, 104 of them used the textual cue given by the cartoon. Therefore, the frequency of three past tense usages was recalculated (see Table 35).

Table 35. *Past Tense Usage on the Pretest and Posttest*

| Total | Pretest | | Posttest | |
|---|---|---|---|---|
| | Count | (%) | Count | (%) |
| Simple past | 1,023 | 97 | 1,115 | 93 |
| Past continuous | 20 | 1 | 66 | 5 |
| Past perfect | 7 | < 1 | 18 | < 1 |
| Total usage | 1,050 | 100 | 1199 | 100 |

*Note*. Past continuous usages based on textual cue given by cartoon were excluded.

Table 36 presents the frequency of the three past tense forms used by the comparison, FFI, and FFI + PF groups to indicate that there were no substantial group differences. For the simple past form, the comparison group produced 287 instances on the pretest and 301 on the posttest, the FFI group produced 376 instances on the pretest and 397 on the posttest, and the FFI + PF group produced 360 instances on the pretest and 417 on the posttest. For the past continuous form, the comparison group produced 36 instances on the pretest and 44 on the posttest, the FFI group produced 48 instances on the pretest and 73 on the posttest, and the FFI + PF group produced 38 instances on the pretest and 54 on the posttest. For the past perfect tense, the comparison group produced four instances on the pretest and five on the posttest, the FFI group produced two

instances on the pretest and seven on posttest, and the FFI + PF group produced one instance on the pretest and six instances on the posttest.

Table 36. *Past Tense Usage by the Three Groups on Pretest and Posttest*

| | Comparison group | | FFI group | | FFI + PF group | |
|---|---|---|---|---|---|---|
| | Pretest | Posttest (gain) | Pretest | Posttest | Pretest | Posttest (gain) |
| Simple Past | 287 | 301 (14) | 376 | 397(19) | 360 | 417 (57) |
| Past continuous | 36 | 44 (8) | 48 | 73 (25) | 38 | 54 (16) |
| Past perfect | 4 | 5 (1) | 2 | 7 (5) | 1 | 6 (5) |
| Total | 327 | 350 (23) | 426 | 477 (51) | 399 | 477 (78) |

*Note*. Past continuous usages based on textual cue given by cartoon are included above. FFI = Form-focused instruction; FFI + PF = Form-focused instruction + peer feedback.

## Summary

First, I first presented the preliminary analyses of the data collected from the English Speaking Learner Affect questionnaire and Self-Assessment of Speaking Task questionnaire using WINSTEPS. The Rasch analyses showed that the measures in both questionnaires were unidimensional; however, several items were deleted from each questionnaire. Second, I presented the preliminary analysis of the raters using FACETS. The multifaceted Rasch analysis of the ratings showed that the rating results were consistent. Third, I presented a discussion of the three past tense forms observed. In the next chapter, I present the results that address each hypothesis.

# CHAPTER 5

# RESULTS

The purpose of this chapter is to present the quantitative results of the six CAF measurements, rater assessments, and questionnaire measures as well as descriptive details of the weekly changes of nine participants selected for qualitative interpretations. The results are organized according to each of the five research hypotheses presented in Chapter 2.

## Research Hypothesis 1: Within-Group Speaking Development

Hypothesis 1a stated that participants who received form-focused instruction (FFI group) and participants who received form-focused instruction and peer feedback (FFI + PF) would improve syntactic accuracy because form-focused instruction explicitly directed them to pay attention to form, especially the use of past tense forms (DeKeyser, 1998; Goh & Burns, 2012; Norris & Ortega, 2000; Spada & Tomita, 2010). Participants in the intervention groups noticed the target forms (Schmidt, 1990, 2001), consciously compared their own output and the target forms (Schmidt & Frota, 1986), and modified their output (Swain, 1985). The participants in the comparison group who did not receive form-focused instruction would not improve syntactic accuracy because they were not explicitly directed to pay attention to the target forms. Hypothesis 1b stated that all three groups would improve syntactic complexity and oral fluency because task repetition has been found to lead to such improvements (Ahmadian, 2011; Bygate, 2001).

In order to answer Hypotheses 1a and 1b, three sets of one-way repeated-measures ANOVAs were conducted to investigate within-group changes for the three groups. The factor was time (pretest and posttest) and the dependent variables were the six CAF measurements: mean length of T-unit, clause/T-unit ratio, percentage of error free T-units, percentage of accurate past tense verb usage, speech rate, and self-repair.

The descriptive statistics for the six variables at pretest and posttest are presented for the three groups in Table 37. Overall, the level of increases and decreases on the posttest compared to the pretest for all three groups were similar except for the percentage of error-free T-units and the percentage of accurate past tense verb usage. The FFI and FFI + PF groups made large increases in these two variables. All three groups made almost no change in the mean length of T-unit and clause/T-unit ratio. In addition, all groups slightly increased in speech rate and slightly decreased in self-repair. For the FFI group, five out of the six variables improved on the posttest compared to the pretest. The percentage of error-free T-units on the posttest increased by 46.40% and the percentage of accurate past tense verb usage increased by 56.51%. For the FFI + PF group, five out of the six variables improved on the posttest compared to the pretest. The percentage of error-free T-units increased by 42.23% and the percentage of accurate past tense verb usage increased by 44.04%. For the comparison group, both the percentage of error-free T-units and the percentage of accurate past tense verb increased by 0.02%.

Table 37. *Means (SD) for Pretest and Posttest Performance of the Three Groups*

| | Comparison group | | Form-focused Instruction group | | Form-focused Instruction + Peer feedback group | |
|---|---|---|---|---|---|---|
| | Pretest | Posttest | Pretest | Posttest | Pretest | Posttest |
| Mean length of T-unit | 9.11 (1.52) | 9.81 (1.54) | 8.72 (1.45) | 8.79 (1.23) | 9.57 (1.42) | 9.84 (1.60) |
| Clause/T-unit ratio | 1.19 (0.10) | 1.24 (0.12) | 1.19 (0.09) | 1.22 (0.12) | 1.25 (0.14) | 1.26 (0.10) |
| Error-free T-units (%) | 52.13 (18.77) | 53.17 (13.35) | 43.58 (23.67) | 63.80 (12.68) | 50.36 (18.54) | 71.63 (14.22) |
| Accurate tense (%) | 68.56 (15.32) | 69.76 (22.15) | 54.05 (29.96) | 84.59 (13.10) | 60.40 (23.98) | 87.00 (13.48) |
| Speech rate | 84.80 (16.17) | 91.45 (14.79) | 71.36 (15.69) | 76.69 (10.92) | 80.96 (16.61) | 85.21 (14.19) |
| Self-repair | 7.92 (3.95) | 9.44 (5.20) | 6.85 (4.59) | 7.39 (5.41) | 7.03 (5.02) | 7.45 (4.72) |

*Note*. A higher score in self-repair indicates decrease in oral fluency.

Tables 38 to 43 present the assumption of univariate normality for the six measurements—mean length of T-unit, clause/T-unit ratio, percentage of error free T-units, percentage of accurate past tense verb usage, speech rate, and self-repair—at the pretest and posttest. The raw scores of all six measures were converted into *z*-scores; there were no univariate outliers above ±3.29 (Field, 2009). Skewness and kurtosis statistics were calculated and the assumption of univariate normality for all six measures was met because they were all below ±2.00 (George & Mallery, 2016).

Table 38. *Descriptive Statistics for Mean Length of T-Unit*

| | Pretest | Posttest | Gain score |
|---|---|---|---|
| *M* | 9.12 | 9.42 | 0.30 |
| *SE* | 0.15 | 0.15 | 0.13 |
| 95% CI | [8.83, 9.41] | [9.12, 9.71] | [8.52, 9.26] |
| *SD* | 1.49 | 1.52 | 1.35 |
| Skewness | 0.65 | 0.81 | 0.00 |
| *SES* | 0.24 | 0.24 | 0.24 |
| Kurtosis | 0.46 | 1.49 | -0.55 |
| *SEK* | 0.47 | 0.47 | 0.47 |

*Note*. SES = Standard error skewness; SEK = Standard error kurtosis.

Table 39. *Descriptive Statistics for Clause/T-Unit Ratio*

| | Pretest | Posttest | Gain score |
|---|---|---|---|
| *M* | 0.84 | 0.87 | 0.03 |
| *SE* | 0.05 | 0.08 | 0.02 |
| 95% CI | [0.75, 0.94] | [0.78, 0.97] | [0.78, 0.97] |
| *SD* | 0.47 | 0.48 | 0.15 |
| Skewness | -0.59 | -0.49 | 0.05 |
| *SES* | 0.24 | -0.24 | 0.24 |
| Kurtosis | -1.46 | -1.55 | 0.96 |
| *SEK* | 0.47 | 0.47 | 0.50 |

*Note*. SES = Standard error skewness; SEK = Standard error kurtosis.

Table 40. *Descriptive Statistics for Percentage of Error-Free T-Units*

| | Pretest | Posttest | Gain score |
|---|---|---|---|
| *M* | 48.11 | 64.11 | 15.99 |
| *SE* | 2.05 | 1.48 | 2.18 |
| 95% CI | [44.05, 52.18] | [61.18, 67.03] | [11.67, 20.32] |
| *SD* | 20.91 | 15.05 | 22.26 |
| Skewness | -0.09 | -0.12 | 0.40 |
| *SES* | 0.24 | 0.24 | 0.24 |
| Kurtosis | -0.91 | -0.46 | -0.28 |
| *SEK* | 0.47 | 0.47 | 0.47 |

*Note*. SES = Standard error skewness; SEK = Standard error kurtosis.

Table 41. *Descriptive Statistics for Percentage of Accurate Past Tense Verb Usage*

| | Pretest | Posttest | Gain score |
|---|---|---|---|
| *M* | 60.14 | 81.62 | 21.47 |
| *SE* | 2.61 | 1.53 | 2.89 |
| 95% CI | [54.98, 65.32] | [78.59, 83.64] | [15.75, 27.19] |
| *SD* | 26.58 | 15.56 | 29.43 |
| Skewness | -0.47 | -0.78 | 0.53 |
| *SES* | 0.24 | 0.24 | 0.24 |
| Kurtosis | -0.69 | 0.11 | -0.14 |
| *SEK* | 0.47 | 0.47 | 0.47 |

*Note*. SES = Standard error skewness; SEK = Standard error kurtosis.

Table 42. *Descriptive Statistics for Speech Rate*

| | Pretest | Posttest | Gain score |
|---|---|---|---|
| *M* | 78.10 | 83.35 | 5.25 |
| *SE* | 1.66 | 1.40 | 1.30 |
| 95% CI | [74.80, 81.40] | [80.57, 86.13] | [2.65, 7.85] |
| *SD* | 16.96 | 14.30 | 13.34 |
| Skewness | -0.40 | 0.66 | 0.45 |
| *SES* | 0.24 | 0.24 | 0.24 |
| Kurtosis | -0.13 | 0.83 | 0.60 |
| *SEK* | 0.47 | 0.47 | 0.47 |

*Note*. SES = Standard error skewness; SEK = Standard error kurtosis.

Table 43. *Descriptive Statistics for Self-Repair*

| | Pretest | Posttest | Gain score |
|---|---|---|---|
| *M* | 7.92 | 7.90 | 0.73 |
| *SE* | 0.79 | 0.54 | 0.42 |
| 95% CI | [6.29, 9.55] | [6.90, 8.90] | [-0.10, 1.56] |
| *SD* | 3.95 | 5.14 | 4.26 |
| Skewness | 0.98 | 0.88 | 0.25 |
| *SES* | 0.46 | 0.24 | 0.24 |
| Kurtosis | 0.58 | 0.60 | 0.77 |
| *SEK* | 0.90 | 0.47 | 0.47 |

*Note*. SES = Standard error skewness; SEK = Standard error kurtosis.

For the FFI within-group comparisons, six one-way repeated-measures ANOVAs were conducted for the following six dependent variables: mean length of T-unit (MLT), clause/T-unit ratio, percentage of error-free T-units, percentage of accurate past tense verb usage, speech rate and self-repair respectively (See Table 44). The one-way repeated-measures ANOVA was chosen because there was no need to examine the interaction effects for within-group comparisons. Results for the FFI group indicated a significant effect for percentage of error-free T-units, Wilk's $\Lambda = .51$, $F(1,40) = 37.81$, $p < .01$, partial $\eta^2 = .49$, percentage of accurate past tense verb usage, Wilk's $\Lambda = .49$, $F(1,40) = 39.28$, $p < .01$, partial $\eta^2 = .51$, and speech rate, Wilk's $\Lambda = .81$, $F(1,40) = 9.57$, $p < .01$, partial = .19. On the other hand, there was no significant effect for mean length of T-unit, Wilk's $\Lambda = 1.00$, $F(1,40) = .13$, $p = .72$, partial $\eta^2 = .01$, clause/T-unit ratio, Wilk's $\Lambda = .93$, $F(1,40) = 3.16$, $p = .08$, partial $\eta^2 = .07$, and self-repair, Wilk's $\Lambda = .99$, $F(1,40) = .60$, $p = .44$, partial $\eta^2 = .02$.

Table 44. *Repeated-Measures ANOVA of Six Variables for the FFI Group*

| Significant | | | | Insignificant | | | |
|---|---|---|---|---|---|---|---|
| Measure | *F* | *p* | η² | Measure | *F* | *p* | η² |
| PEFT | 37.81 | < .01 | .49 | MLT | 0.13 | .72 | .01 |
| PAPTU | 39.28 | < .01 | .51 | C/T ratio | 3.16 | .08 | .07 |
| SR | 9.57 | < .01 | .19 | S | 0.60 | .44 | .02 |

*Note.* PEFT = Percentage of error-free T-units, PAPTU = percentage of accurate past tense usage, MLT = Mean length of T-unit, C/T ratio = clause/T-unit ratio, SR = speech rate, and S = self-repair.

For the within-group comparison of the FFI + PF group, six one-way repeated-measures ANOVAs were conducted for the following six dependent variables: mean length of T-unit (MLT), clause/T-unit ratio, percentage of error-free T-units, percentage of accurate past tense verb usage, speech rate and self-repair (See Table 45). The results indicated a significant effect for the percentage of error-free T-units, Wilk's $\Lambda = .53$, $F(1,37) = 32.80$, $p < .01$, partial $\eta^2 = .47$, and the percentage of accurate past tense verb usage, Wilk's $\Lambda = .47$, $F(1,40) = 41.97$, $p < .01$, partial $\eta^2 = .53$. On the other hand, there were no significant effects for mean length of T-unit, Wilk's $\Lambda = .96$, $F(1,37) = 1.40$, $p = .26$, partial $\eta^2 = .04$, clause/T-unit ratio, Wilk's $\Lambda = 1.00$, $F(1,37) = .06$, $p = .81$, partial $\eta^2 = < .01$, speech rate, Wilk's $\Lambda = .99$, $F(1,37) = .36$, $p = .55$, partial $\eta^2 = .01$, and self-repair, Wilk's $\Lambda = .93$, $F(1,37) = 3.00$, $p = .09$, partial $\eta^2 = .08$.

Table 45. *Repeated-Measures ANOVA of Six Variables for the FFI + PF Group*

| Significant | | | | Insignificant | | | |
|---|---|---|---|---|---|---|---|
| Measure | *F* | *p* | η² | Measure | *F* | *p* | η² |
| PEFT | 32.80 | < .01 | .47 | MLT | 1.40 | .26 | .04 |
| PAPTU | 41.97 | < .01 | .53 | C/T ratio | 0.06 | .81 | < .01 |
| | | | | SR | 0.36 | .55 | .01 |
| | | | | S | 3.00 | .09 | .08 |

*Note.* PEFT = Percentage of error-free T-units, PAPTU = percentage of accurate past tense usage, MLT = Mean length of T-unit, C/T ratio = clause/T-unit ratio, SR = speech rate, and S = self-repair.

For the within-group comparison of the comparison group, six one-way repeated-measures ANOVAs were conducted for the following six dependent variables: mean length of T-unit (MLT), clause/T-unit ratio, percentage of error-free T-units, percentage of accurate past tense verb usage, speech rate and self-repair (See Table 46). The results indicated a significant effect for mean length of T-unit, Wilk's $\Lambda = .75$, $F(1,24) = 8.22$, $p < .01$, partial $\eta^2 = .26$, and speech rate, Wilk's $\Lambda = .82$, $F(1,24) = 5.40$, $p = .03$, partial $\eta^2 = .18$. On the other hand, there was no significant effect for the clause/T-unit ratio, Wilk's $\Lambda = .92$, $F(1,24) = 2.10$, $p = .16$, partial $\eta^2 = .08$, percentage of error-free T-units, Wilk's $\Lambda = 1.00$, $F(1,24) = .10$, $p = .76$, partial $\eta^2 < .01$, percentage of accurate past tense verb usage, Wilk's $\Lambda = .99$, $F(1,24) = .08$, $p = .79$, partial $\eta^2 < .01$, and self-repair, Wilk's $\Lambda = .87$, $F(1,24) = 3.67$, $p = .07$, partial $\eta^2 = .13$.

Table 46. *Repeated-Measures ANOVA of Six Variables for the Comparison Group*

| Significant | | | | Insignificant | | | |
|---|---|---|---|---|---|---|---|
| Measure | *F* | *p* | $\eta^2$ | Measure | *F* | *p* | $\eta^2$ |
| MLT | 8.22 | < .01 | .26 | C/T ratio | 2.10 | .16 | .08 |
| SR | 5.40 | .03 | .18 | PAPTU | 0.08 | .79 | .01 |
| | | | | PEFT | 0.10 | .76 | < .01 |
| | | | | S | 3.67 | .07 | .13 |

*Note.* PEFT = Percentage of error-free T-units, PAPTU = percentage of accurate past tense usage, MLT = Mean length of T-unit, C/T ratio = clause/T-unit ratio, SR = speech rate, and S = self-repair.

## Research Hypothesis 2: Effects of FFI on Speaking Development

Hypothesis 2a stated that the FFI group would improve syntactic accuracy more than the comparison group because form-focused instruction explicitly directed them to pay attention to form, especially the use of past tense forms. FFI group participants noticed the target forms (Schmidt, 1990, 2001), consciously compared their own output

and the target forms (Schmidt & Frota, 1986), and modified their output (Swain, 1985). Task repetition alone did not explicitly direct the comparison group participants to pay attention to the target forms and notice errors in their use of the target forms. The FFI + PF group should improve syntactic accuracy more than the FFI group because more explicit form-focused instruction should lead to more syntactic accuracy development (DeKeyser, 1998; Goh & Burns, 2012; Norris & Ortega, 2000; Spada & Tomita, 2010). Hypothesis 2b stated that the FFI + PF and the FFI groups would not show greater improvements in either syntactic complexity or oral fluency than the comparison group because the FFI treatments do not target these two variables.

Before the multivariate and univariate analyses were conducted, differences between the three groups were examined by comparing the mean gain scores for the six variables (posttest - pretest) of the three groups (Table 47).

Table 47. *Gain Score Means (SD) for the Six Variables*

| | Comparison | FFI | FFI + PF |
|---|---|---|---|
| | *M* (*SD*) | *M* (*SD*) | *M* (*SD*) |
| Mean length of T-unit | 0.69 (1.20) | 0.08 (1.35) | 0.27 (1.43) |
| Clause/T-unit ratio | 0.04 (0.17) | 0.04 (0.13) | 0.01 (0.16) |
| Error-free T-units (%) | 1.03 (16.48) | 20.23 (21.06) | 21.27 (22.90) |
| Accurate tense (%) | -1.19 (21.92) | 30.54 (30.31) | 26.60 (25.31) |
| Speech rate | 6.64 (14.31) | 5.32 (11.02) | 4.25 (15.14) |
| Self-repair | 1.52 (3.96) | 0.54 (4.42) | 0.42 (4.33) |

*Note*. Gain store in self-repair indicates decrease in performance. FFI = form-focused instruction; FFI + PF = form-focused instruction and peer feedback.

Table 47 indicates that the FFI group and FFI + PF group outperformed the comparison group in the percentage of error-free T-units and percentage of accurate past tense verb usage. The comparison group slightly outperformed the FFI and FFI +

PF groups in the mean length of T-unit. There were very small differences between the FFI group and the FFI + PF group.

Pearson correlation coefficients were examined between the six variables at the pretest and posttest respectively to assess multicollinearity (Tabachnick & Fidell, 2007). Table 48 shows the correlation coefficients for the six variables at the pretest for all participants. There were two high correlation coefficients, the percentage of error-free T-units and the percentage of accurate past tense verb usage correlated at $r = .87$, $p < .01$, and speech rate and self-repairs correlated at $r = .84$, $p < .01$. While these two correlations are considered to be large because they are above .50 (Green & Salkind, 2011), they are also considered to be acceptable and do not cause concern for multicollinearity as they are under the benchmark < .90 (Tabachnick & Fidell, 2007).

Table 48. *Correlations Between Six Variables on the Pretest*

| | | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| 1. | MLT | — | | | | | |
| 2. | C/T ratio | -.19 | — | | | | |
| 3. | PEFT | .17 | -.08 | — | | | |
| 4. | PAPTU | .18 | -.01 | .87** | — | | |
| 5. | SR | .13 | -.14 | -.03 | -.01 | — | |
| 6. | Self-repair | .19 | .15 | -.03 | -.01 | .84** | — |

*Note.* Correlation is significant at < .01 (2-tailed). MLT = Mean length of T-unit, C/T ratio = clause/T-unit ratio, PEFT = Percentage of error-free T-units, PAPTU = percentage of accurate past tense usage, SR = speech rate.

Table 49 shows the correlation coefficients for the six variables at the posttest for all participants. There was one high correlation coefficient; the percentage of error-free T-units and the percentage of accurate past tense verb usage correlated at $r = .64$, $p < .01$. there were four medium correlation coefficients, clause/T-unit ratio and

percentage of error-free T-units correlated at $r = -.40$, $p < .01$, clause/T-unit ratio and percentage accurate past tense verb usage correlated at $r = -.24$, $p < .01$, percentage of error-free T-units and self-repair correlated at $r = -.26$, $p < .01$, and speech rate and self-repair correlated at $r = -.29$, $p < .01$. All correlation coefficients at the posttest were acceptable because they are under the maximum correlation coefficient benchmark of < .90 (Tabachnick & Fidell, 2007).

Table 49. *Correlations Between Six Variables on the Posttest*

| | | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| 1. | MLT | — | | | | | |
| 2. | C/T ratio | -.17 | — | | | | |
| 3. | PEFT | .09 | -.40** | — | | | |
| 4. | PAPPU | .18 | -.24** | .64** | — | | |
| 5. | SR | .15 | -.09 | .89 | .03 | — | |
| 6. | Self-repair | .09 | .03 | -.26** | -.13 | -.29** | — |

*Note.* **Correlation is significant at < .01 (2-tailed). MLT = Mean length of T-unit, C/T ratio = clause/T-unit ratio, PEFT = Percentage of error-free T-units, PAPTU = percentage of accurate past tense usage, SR = speech rate.

A one-way MANOVA was conducted to determine the effect of form-focused instruction using six dependent variables: mean length of T-unit, clause/T-unit ratio, percentage of error free T-units, percentage of accurate past tense verb usage, speech rate, and self-repair. A MANOVA was used because the correlation coefficients between the six variables were under the < .90 benchmark for multicollinearity (Field, 2009; Tabachnick & Fidell, 2007). The independent variable was group (three levels: comparison, FFI, and FFI + PF). The assumptions of the analysis were checked and met. A significant difference was found among the three groups on the dependent measures, Wilk's $\Lambda = .72$, $F(12,192) = 2.84$, $p = < .01$, partial $\eta^2 = .15$.

The MANOVA was followed up with six one-way ANOVAs for the six dependent variables (See Table 50). A Holm's Bonferroni adjustment was made to control for Type I errors. Using a Bonferroni adjustment, each ANOVA was tested at the $p = .008$. Among the six ANOVAs conducted, two were significant. The ANOVA of percentage of error-free T-units was significant, $F(2,101) = 8.55$, $p = .001$, partial $\eta^2 = .15$. A follow-up post hoc analysis was conducted for the percentage of error-free T-units. Because the sample sizes differed, Hochberg's GT2 post hoc test was used to investigate pairwise differences. The results indicated that the FFI group significantly outperformed the comparison group, $p = .05$, and the FFI + PF outperformed the comparison group, $p = .05$. The FFI + PF group slightly outperformed the FFI group, but it was not significant, $p = .99$. The ANOVA for the percentage of accurate past tense verb usage was also significant, $F(2,101) = 12.55$, $p < .001$, partial $\eta^2 = .20$. The follow-up post hoc analysis indicated that the FFI group significantly outperformed the comparison group, $p = .05$, and the FFI + PF group outperformed the comparison group, $p = .05$. The FFI + PF group slightly outperformed the FFI group, but it was not significant, $p = .88$.

The ANOVAs for mean length of T-unit, clause/T-unit ratio, speech rate, and self-repair were not significant. The ANOVA for mean length of T-unit was $F(2,101) = 1.63$, $p = .20$, partial $\eta^2 = .20$, clause/T-unit ratio was $F(2,101) = .67$, $p = .51$, partial $\eta^2 = .01$, speech rate was $F(2,101) = .24$, $p = .79$, partial $\eta^2 = .01$, and self-repair was $F(2,101) = .57$, $p = .57$, partial $\eta^2 = .01$.

Table 50. *Univariate Analyses of Variance for the Six Variables*

| Variable | *F* | *p* | η² |
|---|---|---|---|
| Mean length of T-unit | 1.63 | .20 | .20 |
| Clause/T-unit ratio | 0.67 | .51 | .01 |
| Percentage of error-free T-units | 8.55 | .001 | .15 |
| Percentage of accurate past tense usage | 12.55 | .001 | .20 |
| Speech rate | 0.24 | .79 | .01 |
| Self-repair | 0.57 | .57 | .01 |

Analytical ratings were also used to examine Hypothesis 2, which stated that the FFI + PF group participants who received form-focused instruction and peer feedback would significantly improve more than the FFI group participants. In addition, both the FFI + PF and FFI group participants would improve significantly more than the comparison group participants. Table 51 presents the descriptive statistics for the gains in analytical rating (posttest - pretest) based on the Rasch fair measure average. The descriptive statistics indicate that the FFI + PF group improved more than the FFI and the comparison groups. The FFI and the comparison groups made almost no change.

Table 51. *Descriptive Statistics of Analytically Rated Gain of the Three Groups*

| | Comparison | FFI | FFI + PF |
|---|---|---|---|
| *M* | -0.05 | < 0.01 | 0.30 |
| *SE* | 0.13 | 0.10 | 0.11 |
| *95%CI* | [-0.31, 0.22] | [-0.21, 0.22] | [0.09, 0.53] |
| *SD* | 0.64 | 0.66 | 0.65 |
| *SK* | -0.53 | -0.91 | 0.41 |
| *SES* | 0.46 | 0.37 | 0.38 |
| *KT* | -0.01 | 3.60 | -0.13 |
| *SEK* | 0.90 | 0.72 | 0.70 |

*Note*. CI = 95% confidence interval; SK = skewness; KT = kurtosis; SES = Standard error skewness; SEK = Standard error kurtosis. All statistics are based on Rasch logits.

A one-way ANOVA was conducted to evaluate the effects of form-focused instruction on the rated speaking performance. Group (three levels: comparison, FFI,

and FFI + PF) was the independent variable and gains in fair average logits was the dependent variable. The result was not significant, $F(2,101) = 2.87$, $p = .06$, partial $\eta^2 = .05$. The effect size, as assessed by partial $\eta^2$, was weak, with the effects of form-focused instruction accounting for 5.00% of the variance in the dependent variable.

### Research Hypothesis 3: Effects of FFI on Trajectory of Changes

Hypothesis 3 stated that the FFI + PF group would improve syntactic accuracy more quickly than the FFI group and the comparison, and the FFI group would improve more quickly than the comparison group because FFI + PF group received most form-focused instruction and form-focused instruction increases syntactic accuracy (Ahmadian & Tavakoli, 2011; Mochizuki & Ortega, 2008; Williams, 2005; Yuan & Ellis, 2003). In order to examine this hypothesis, the week-to-week trajectory of changes in the six CAF measurements (mean length of T-unit, clauses/T-unit ratio, percentage of error-free T-units, percentage of accurate past tense verb usage, speech rate, and self-repairs) of nine representative participants were examined. Three representative participants from each group were chosen. The three comparison group participants examined were Yuka, Yuu, and Yusuke; The three FFI group participants examined were Daiki, Daiichi, and Daisuke; and the three FFI + PF group participants examined were Mana, Manabu, and Musashi. Weekly changes of the six measurements are displayed in graphs and tables to show trends over time between and within groups.

**Changes in Mean Length of T-Unit**

Figure 11 shows the changes in mean length of T-unit of the nine representative participants from the three groups over time. Overall, no group performed better than the other two groups as all participants' performance of mean length of T-unit fluctuated without noticeable trends. Table 52 presents the descriptive statistics of changes in mean length of T-unit of the nine participants from the three groups from Week 1 to Week 9.

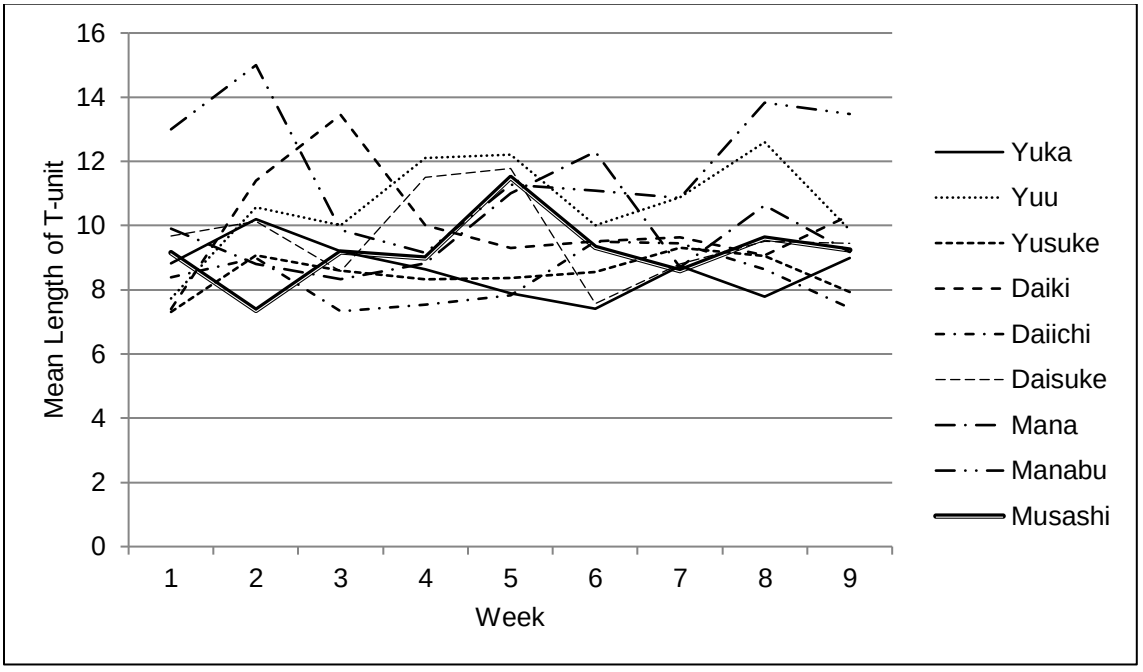



*Figure 11*. Weekly changes in mean length of T-unit. Solid lines are comparison group representatives; Dashed lines are FFI group representatives; Dotted lines are FFI + PF group representatives.

Table 52 shows the descriptive statistics of changes in mean length of T-unit of the nine participants from the three groups from Week 1 to Week 9. The table shows that seven out of the nine participants increased from Week 1 to Week 2 between 4.75%

to 54.05%. Two FFI + PF group participants decreased in mean length of T-unit from Week 1 to Week 2; Mana produced 8.80 MLT, a decrease of 11.11%, and Musashi produced 9.15 MLT, a decrease of 19.56%. Seven participants decreased in mean length of T-unit from Week 2 to Week 3, and their level of decrease ranged from 5.29% to 34.20%. While seven participants decreased in performance from Week 2 to Week 3, two participants increased: Daiki (FFI group) produced 13.45 MLT, a 17.98% increase, and Musashi (FFI + PF group) produced 9.18 MLT, a 24.73% increase. From Week 4 to Week 9, the nine participants continued their fluctuations in mean length of T-unit with most increases and decreases ranging between 10.00-20.00%.

Table 52. *Weekly Performance of Mean Length of T-Unit*

| | Comparison group | | | FFI group | | | FFI + PF group | | |
|---|---|---|---|---|---|---|---|---|---|
| Week | Yuka | Yuu | Yusuke | Daiki | Daiichi | Daisuke | Mana | Manabu | Musashi |
| 1 | 8.83 | 7.73 | 7.31 | 7.40 | 8.38 | 9.67 | 9.90 | 13.00 | 9.15 |
| 2 | 10.20 | 10.56 | 9.07 | 11.40 | 9.00 | 10.13 | 8.80 | 15.00 | 7.36 |
| 3 | 9.20 | 10.00 | 8.59 | 13.45 | 7.33 | 8.58 | 8.33 | 9.87 | 9.18 |
| 4 | 8.64 | 12.11 | 8.33 | 10.00 | 7.54 | 11.50 | 8.82 | 9.15 | 9.00 |
| 5 | 7.88 | 12.22 | 8.37 | 9.31 | 7.82 | 11.78 | 11.00 | 11.27 | 11.50 |
| 6 | 7.41 | 10.00 | 8.56 | 9.50 | 9.50 | 7.57 | 12.30 | 11.08 | 9.33 |
| 7 | 8.75 | 10.88 | 9.31 | 9.64 | 9.45 | 8.80 | 8.69 | 10.85 | 8.60 |
| 8 | 7.78 | 12.60 | 9.06 | 9.07 | 8.64 | 9.50 | 10.64 | 13.82 | 9.62 |
| 9 | 9.00 | 9.86 | 7.94 | 10.33 | 7.44 | 9.45 | 9.17 | 13.47 | 9.25 |

*Note*. FFI = Form-focused instruction; FFI + PF = Form-focused instruction and peer feedback.

**Changes in Clause/T-Unit Ratio**

Figure 12 shows the changes in clause/T-unit ratio of the nine representative participants over time. No group improved more than any other group. The performance of the nine participants fluctuated over time, but the week-to-week fluctuations were smaller compared to the fluctuations in the other measurements.

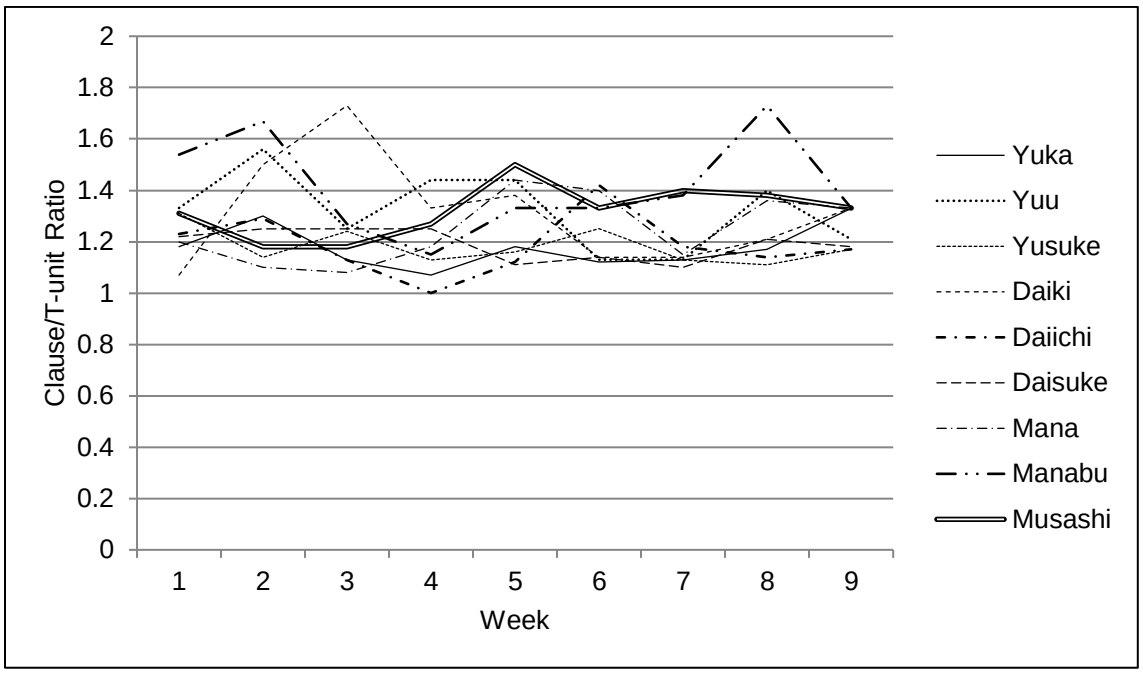



*Figure 12*. Weekly changes in clause/T-unit ratio. Solid lines are Comparison group representatives; Dashed lines are FFI group representatives; Dotted lines are FFI + PF group representatives.

Table 53 presents the descriptive statistics of changes in clause/T-unit ratio of the nine participants of the three groups from Week 1 to Week 9. The table shows that six out of the nine participants increased from Week 1 to Week 2 and the level of increase ranged between 2.46% to 40.19%. Two FFI + PF group participants and one comparison group participant decreased in clause/T-unit ratio in Week 2, Mana produced 1.10 C/T ratio, a decrease of 8.33%, Musashi produced 1.18 C/T ratio, a decrease of 9.92%, and Yusuke produced 1.14 C/T ratio, a decrease of 12.98%. In Week 3, five participants decreased, two participants made no change, and two participants increased in clause/T-unit ratio. From Week 4 to Week 9, the nine participants continued their fluctuations and mostly produced between 1.2 to 1.4 clause/T-unit ratio each week.

Table 53. *Weekly Performance of Clause/T-Unit Ratio*

| | Comparison group | | | FFI group | | | FFI + PF group | | |
|---|---|---|---|---|---|---|---|---|---|
| Week | Yuka | Yuu | Yusuke | Daiki | Daiichi | Daisuke | Mana | Manabu | Musashi |
| 1 | 1.18 | 1.33 | 1.31 | 1.07 | 1.23 | 1.22 | 1.20 | 1.54 | 1.31 |
| 2 | 1.30 | 1.56 | 1.14 | 1.50 | 1.29 | 1.25 | 1.10 | 1.67 | 1.18 |
| 3 | 1.13 | 1.25 | 1.24 | 1.73 | 1.13 | 1.25 | 1.08 | 1.27 | 1.18 |
| 4 | 1.07 | 1.44 | 1.13 | 1.33 | 1.00 | 1.25 | 1.18 | 1.15 | 1.27 |
| 5 | 1.18 | 1.44 | 1.16 | 1.38 | 1.12 | 1.11 | 1.44 | 1.33 | 1.50 |
| 6 | 1.12 | 1.13 | 1.25 | 1.14 | 1.42 | 1.14 | 1.40 | 1.33 | 1.33 |
| 7 | 1.13 | 1.13 | 1.13 | 1.14 | 1.18 | 1.10 | 1.15 | 1.38 | 1.40 |
| 8 | 1.17 | 1.40 | 1.11 | 1.21 | 1.14 | 1.21 | 1.36 | 1.73 | 1.38 |
| 9 | 1.33 | 1.21 | 1.17 | 1.33 | 1.17 | 1.18 | 1.33 | 1.33 | 1.33 |

*Note*. FFI = Form-focused instruction; FFI + PF = Form-focused instruction and peer feedback.

**Changes in the Percentage of Error-Free T-Units**

Figure 13 shows five trends in the percentage of error-free T-units produced by the nine participants over time. First, the FFI + PF group participants outperformed the comparison group and the FFI group participants, and the FFI group participants outperformed the comparison group participants in most weeks. Second, the fluctuations in the percentage of error-free T-units produced by the participants occurred most rapidly in the initial weeks of the research period. From Week 1 to Week 2, all three comparison group participants decreased rapidly in the percentage of error-free T-units, within which one comparison group participant produced no error-free T-unit, whereas some FFI + PF and FFI group participants increased the number of error-free T-units. From Week 2 to Week 3, eight out of nine participants made large increases in the percentage of error-free T-units. Third, the FFI + PF and FFI group participants improved in the percentage of error-free T-units throughout the research period notwithstanding the fluctuations they showed. Fourth, the FFI and FFI + PF group participants continued to improve with fluctuations until later weeks of the

research period, and then the increases flattened off. Finally, the comparison group participants fluctuated in their percentage of error-free T-units throughout the study.

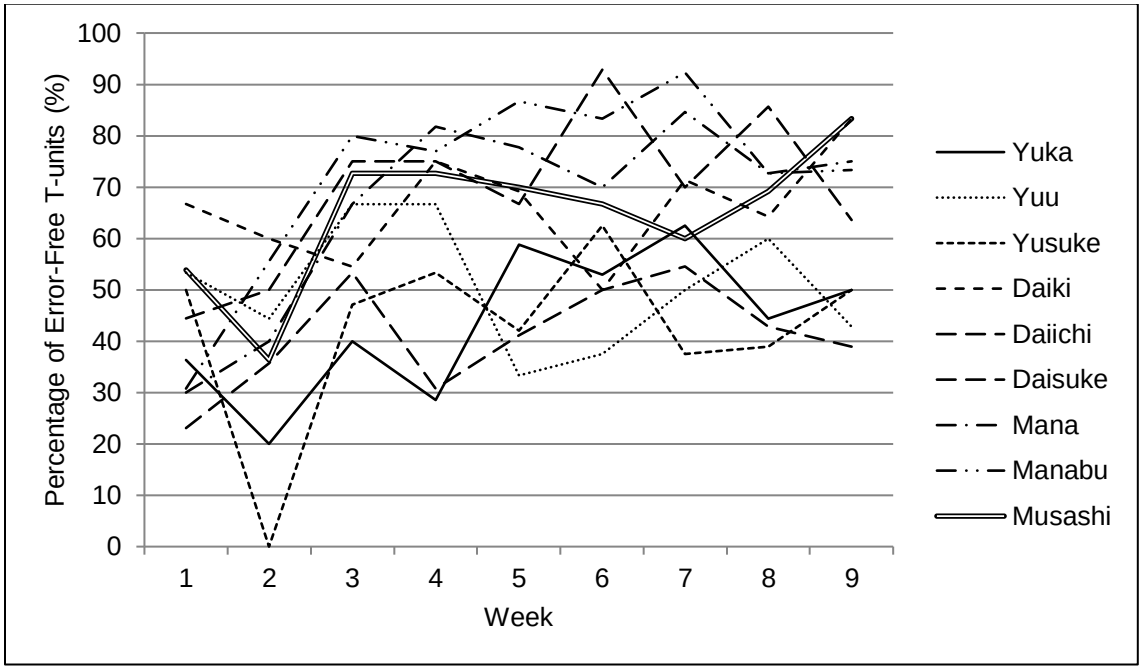



*Figure 13*. Weekly changes in the percentage of error-free T-units. Solid lines are Comparison group representatives; Dashed lines are FFI group representatives; Dotted lines are FFI + PF group representatives.

Table 54 shows the descriptive statistics of changes in the percentage of error-free T-units of the nine participants from Week 1 to Week 9. All three comparison group participants and one FFI + PF group participant decreased in the percentage of error-free T-units from Week 1 to Week 2. Among the four participants who decreased in Week 2, Yuka decreased 16.36% error-free T-units, a decrease of 44.99%, Yuu decreased 8.89% error-free T-units, a decrease of 16.67%, Yusuke decreased 50.00% error-free T-units, a decrease of 100.00%, and Musashi decreased 17.49% error-free T-units, a decrease of 32.48%. In Week 2, Yusuke (comparison group) did not produce

any error-free T-units. In Week 3, eight participants increased the percentage of error-free T-units and the increases ranged from 43.99% to 100.00%. In Week 3, Yuka (comparison group) produced 40.00% error-free T-units, which was double the number of error-free T-units she produced in Week 2 (20.00%), and Musashi (FFI + PF group) produced 72.75% error-free T-units, which was more than double what he produced in Week 2 (36.36%). In Week 3, Daiki (FFI group) was the only participant who decreased in the percentage of error-free T-units; he produced 54.55% error-free T-units, a decrease of 9.08%.

Table 54. *Weekly Performance of the Percentage of Error-Free T-Units*

| | Comparison group | | | FFI group | | | FFI + PF group | | |
|---|---|---|---|---|---|---|---|---|---|
| Week | Yuka | Yuu | Yusuke | Daiki | Daiichi | Daisuke | Mana | Manabu | Musashi |
| 1 | 36.36 | 53.33 | 50.0 | 66.67 | 23.08 | 44.44 | 30.00 | 30.77 | 53.85 |
| 2 | 20.00 | 44.44 | 0.00 | 60.00 | 35.71 | 50.00 | 40.00 | 55.56 | 36.36 |
| 3 | 40.00 | 66.67 | 47.06 | 54.55 | 53.33 | 75.00 | 66.67 | 80.00 | 72.73 |
| 4 | 28.57 | 66.67 | 53.33 | 75.00 | 30.77 | 75.00 | 81.82 | 76.92 | 72.73 |
| 5 | 58.82 | 33.33 | 42.11 | 69.23 | 41.18 | 66.67 | 77.78 | 86.67 | 70.00 |
| 6 | 52.94 | 37.50 | 62.50 | 50.00 | 50.00 | 92.86 | 70.00 | 83.33 | 66.67 |
| 7 | 62.50 | 50.00 | 37.50 | 71.43 | 54.55 | 70.00 | 84.62 | 92.31 | 60.00 |
| 8 | 44.44 | 60.00 | 38.89 | 64.29 | 42.86 | 85.71 | 72.73 | 72.73 | 69.23 |
| 9 | 50.00 | 42.86 | 50.00 | 83.33 | 38.89 | 63.64 | 75.00 | 73.33 | 83.33 |

*Note*. FFI = Form-focused instruction; FFI + PF = Form-focused instruction and peer feedback.

**Changes in the Percentage of Accurate Past Tense Verb Usage**

Figure 14 shows changes in the percentage of accurate past tense verb usage produced by the nine participants. There were a five trends. First, FFI + PF and FFI group participants' percentage of accurate past tense verb usage increased while fluctuating over time. The fluctuations were larger in the initial weeks. Second, the FFI + PF group and FFI group participants outperformed the comparison group participants in most weeks. Third, the FFI + PF participants outperformed FFI participants in the

initial weeks of the research period. Fourth, most participants decreased in the percentage of accurate past tense verb usage in Week 2 and then increased in Week 3. Fifth, participants who reached 100.00% in the percentage of accurate past tense verb usage did not maintain their performance and slightly decreased in the following week.

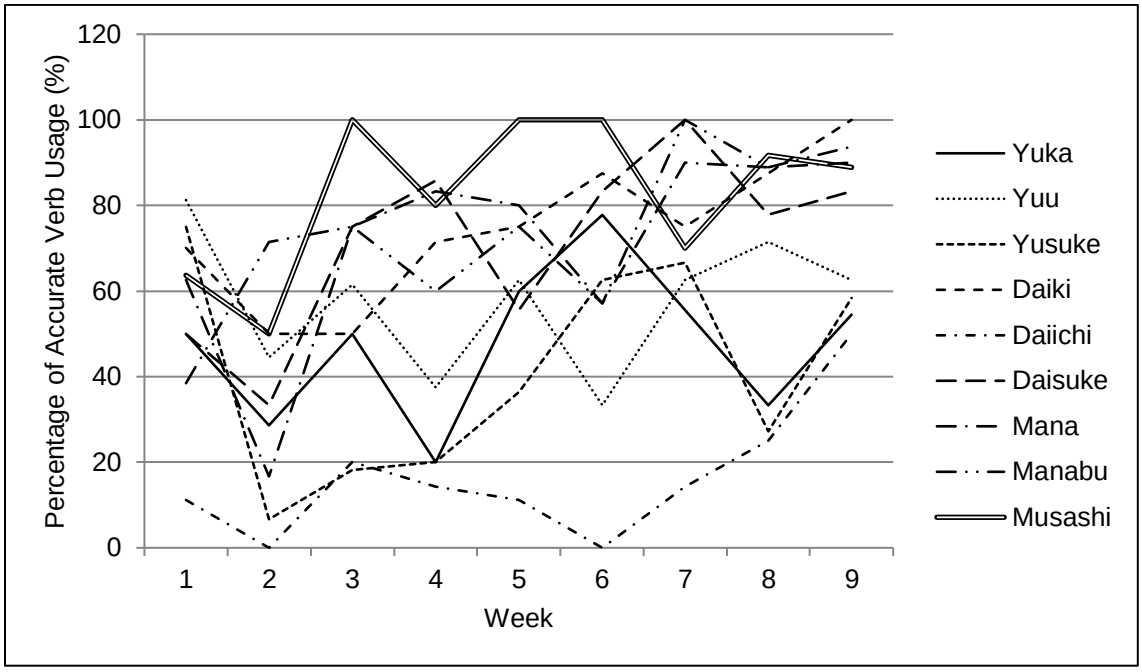



*Figure 14*. Weekly changes in the percentage of accurate past tense verb usage. Solid lines are Comparison group representatives; Dashed lines are FFI group representatives; Dotted lines are FFI + PF group representatives.

Table 55 shows the descriptive statistics of changes in the percentage of accurate past tense usage of the nine participants from the three groups from Week 1 to Week 9. The FFI and FFI + PF group participants increased more in accurate past tense verb usage than the comparison group participants in most weeks. Most participants decreased the percentage of accurate past tense verb usage in Week 2 and then increased in Week 3. In Week 2, eight participants decreased the percentage of accurate

past tense verb usage from 21.43% to 100.00%. In Week 2, the largest decreases were made by Yusuke (comparison group), who produced 6.67% accurate past tense verb usage (which was a 91.11% decrease compared to Week 1) and Daiichi (FFI group), who produced no accurate past tense verbs. On the other hand, Manabu (FFI + PF group) was the only participant who improved in the percentage of accurate past tense verb usage in Week 2, when he produced 71.43%, an 85.73% increase compared to Week 1. In Week 3, all participants increased in the percentage of accurate past tense verb usage from 5.00% to 349.91%. In Week 3, the largest increase was made by Mana (FFI + PF group), who produced 75.00% accurate past tense verb usage, a 349.91% increase compared to Week 2. Daiichi (FFI group) produced 20.00% accurate past tense verb usage, which was a noticeable increase, as he produced no accurate past tense verbs in Week 2. Daiki (FFI group) made no change in percentage accurate past tense verb usage in Week 3 compared to Week 2. Over the nine weeks, Musashi (FFI + PF group) was the highest achiever; he produced 100.00% accurate past tense verb usage in Weeks 3, 5, and 6. Participants who achieved 100.00% accurate past tense verb usage at one point during the research period were unable to maintain their 100.00% and slightly decreased in the following week. Musashi produced 80.00% accurate past tense verb usage in Week 3, which was a decrease from 100.00% in Week 2 (a decrease of 20.00%), and he produced 70.00% accurate past tense verb usage in Week 7, which was a decrease from 100.00% in Week 6 (a decrease of 30.00%). Daisuke produced 77.78% accurate past tense verb usage in Week 8, when he decreased from 100.00% in Week 7 (a decrease of 22.22%). The comparison group participants continued to fluctuate in

their percentage of accurate past tense verb usage throughout the nine weeks; they increased in some weeks and decreased in other weeks with no within-group tendency.

Table 55. *Weekly Performance of the Percentage of Accurate Past Tense Verb Usage*

| | Comparison group | | | FFI group | | | FFI + PF group | | |
|---|---|---|---|---|---|---|---|---|---|
| Week | Yuka | Yuu | Yusuke | Daiki | Daiichi | Daisuke | Mana | Manabu | Musashi |
| 1 | 50.00 | 81.25 | 75.00 | 70.00 | 11.11 | 50.00 | 62.50 | 38.46 | 63.64 |
| 2 | 28.57 | 44.44 | 6.67 | 50.00 | 0.00 | 33.33 | 16.67 | 71.43 | 50.00 |
| 3 | 50.00 | 61.54 | 18.18 | 50.00 | 20.00 | 75.00 | 75.00 | 75.00 | 100.00 |
| 4 | 20.00 | 37.50 | 20.00 | 71.43 | 14.29 | 85.71 | 83.33 | 60.00 | 80.00 |
| 5 | 60.00 | 62.50 | 36.36 | 75.00 | 11.11 | 55.56 | 80.00 | 75.00 | 100.00 |
| 6 | 77.78 | 33.33 | 62.50 | 87.50 | 0.00 | 83.33 | 57.14 | 57.14 | 100.00 |
| 7 | 55.56 | 62.50 | 66.67 | 75.00 | 14.29 | 100.00 | 90.00 | 100.00 | 70.00 |
| 8 | 33.33 | 71.43 | 27.27 | 87.50 | 25.00 | 77.78 | 88.89 | 88.89 | 91.67 |
| 9 | 54.55 | 62.50 | 58.33 | 100.00 | 50.00 | 83.33 | 90.00 | 93.75 | 88.89 |

*Note*. FFI = Form-focused instruction; FFI + PF = Form-focused instruction and peer feedback.

**Changes in Speech Rate**

Figure 15 shows the changes in speech rate of the nine participants over nine weeks. Overall, the groups' performances were similar as the performance of the nine participants fluctuated over time. Nevertheless, three trends were identified. First, the participants had smaller week-to-week fluctuations in speech rate compared with accuracy measurements. Second, the comparison, FFI, and FFI + PF group participants increased in speech rate with fluctuations over the nine weeks whereas the FFI + PF participants had week-to-week fluctuations but made almost no gains after nine weeks. Third, almost all participants slightly decreased in speech rate around Week 7.

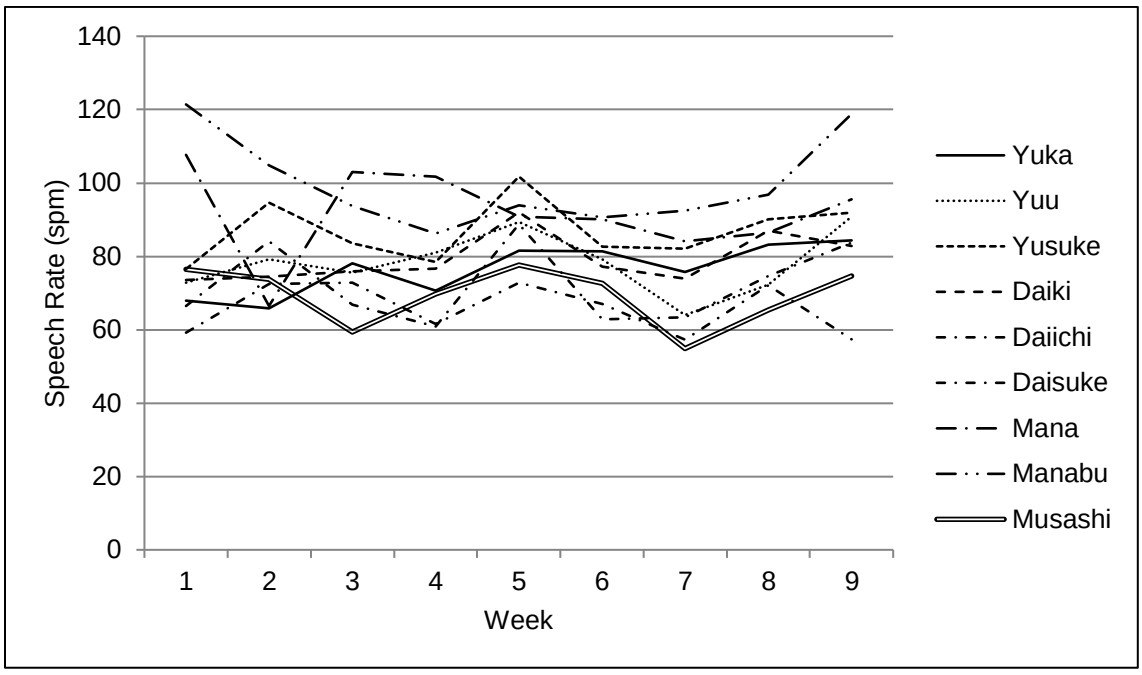



*Figure 15.* Weekly changes in speech rate. Solid lines are the comparison group representatives; Dashed lines are the FFI group representatives; Dotted lines are FFI + PF group representatives.

Table 56 shows the descriptive statistics of changes in speech rate of the nine participants from the three groups from Week 1 to Week 9. In Week 2, all three FFI + PF group participants decreased speech rate whereas all three FFI group and two comparison participants increased speech rate. In Week 2, Mana (FFI + PF group) produced 66.48 spm, a decrease of 38.26%, Manabu (FFI + PF group) produced 104.87 spm, a decrease of 13.63%, and Musashi (FFI + PF group) produced 73.71 spm, a decrease of 3.60%. Second, the comparison and FFI group participants gradually increased in speech rate with fluctuations over the nine weeks whereas the FFI + PF participants had week-to-week fluctuations but made almost no gains throughout the weeks. In Week 7, seven out of the nine participants decreased in speech rate from the previous week from 0.71% to 24.38%. Daiichi (FFI group) and Manabu (FFI + PF

group) increased in speech rate in Week 7, but their increase was very small, Daiichi produced 63.38 spm, an increase of 0.94%, and Manabu produced 92.43 spm, an increase of 1.97%. Finally, three FFI + PF group participants made noticeable increases in speech rate between Week 7 and Week 9. Mana (FFI + PF group) produced 86.43 spm in Week 8, an increase of 2.67%, and then 95.54 spm in Week 9, an increase of 10.54%. Manabu (FFI + PF group) produced 96.81 spm in Week 8, an increase of 4.74%, and then 118.78 spm in Week 9, an increase of 22.69%. Musashi (FFI + PF group) produced 65.38 spm in Week 8, an increase of 19.15%, and then 74.68 spm in Week 9, an increase of 14.22%.

Table 56. *Weekly Performance of Speech Rate*

| | Comparison group | | | FFI group | | | FFI + PF group | | |
|---|---|---|---|---|---|---|---|---|---|
| Week | Yuka | Yuu | Yusuke | Daiki | Daiichi | Daisuke | Mana | Manabu | Musashi |
| 1 | 67.99 | 72.82 | 76.55 | 73.56 | 66.51 | 59.17 | 107.67 | 121.42 | 76.46 |
| 2 | 65.83 | 79.12 | 94.71 | 74.51 | 84.11 | 72.47 | 66.48 | 104.87 | 73.71 |
| 3 | 78.13 | 75.48 | 83.47 | 75.92 | 66.84 | 72.79 | 103.05 | 93.68 | 59.38 |
| 4 | 70.56 | 81.02 | 78.39 | 76.68 | 60.88 | 61.60 | 101.67 | 86.23 | 69.77 |
| 5 | 81.53 | 89.36 | 101.74 | 92.00 | 88.77 | 72.81 | 90.91 | 93.96 | 77.64 |
| 6 | 81.33 | 79.15 | 82.72 | 77.27 | 62.79 | 66.93 | 90.08 | 90.64 | 72.56 |
| 7 | 75.74 | 64.01 | 82.13 | 73.94 | 63.38 | 57.29 | 84.18 | 92.43 | 54.87 |
| 8 | 83.17 | 72.29 | 90.17 | 87.04 | 74.60 | 72.14 | 86.43 | 96.81 | 65.38 |
| 9 | 84.35 | 90.82 | 92.00 | 82.76 | 83.71 | 57.36 | 95.54 | 118.78 | 74.68 |

*Note*. FFI = Form-focused instruction; FFI + PF = Form-focused instruction and peer feedback; The unit of measurement is syllables per minute.

**Changes in Self-Repair**

Figure 16 shows the changes in self-repair for the nine participants over the nine weeks. Three trends can be observed. First, the groups' performances were similar, as the number of self-repairs fluctuated over the nine weeks and there was no consistency. Fluctuations in the number of self-repairs were more rapid compared with fluctuations

of other measurements. A number of participants made large increases throughout the nine weeks and increased more than 100.00% in self-repairs.

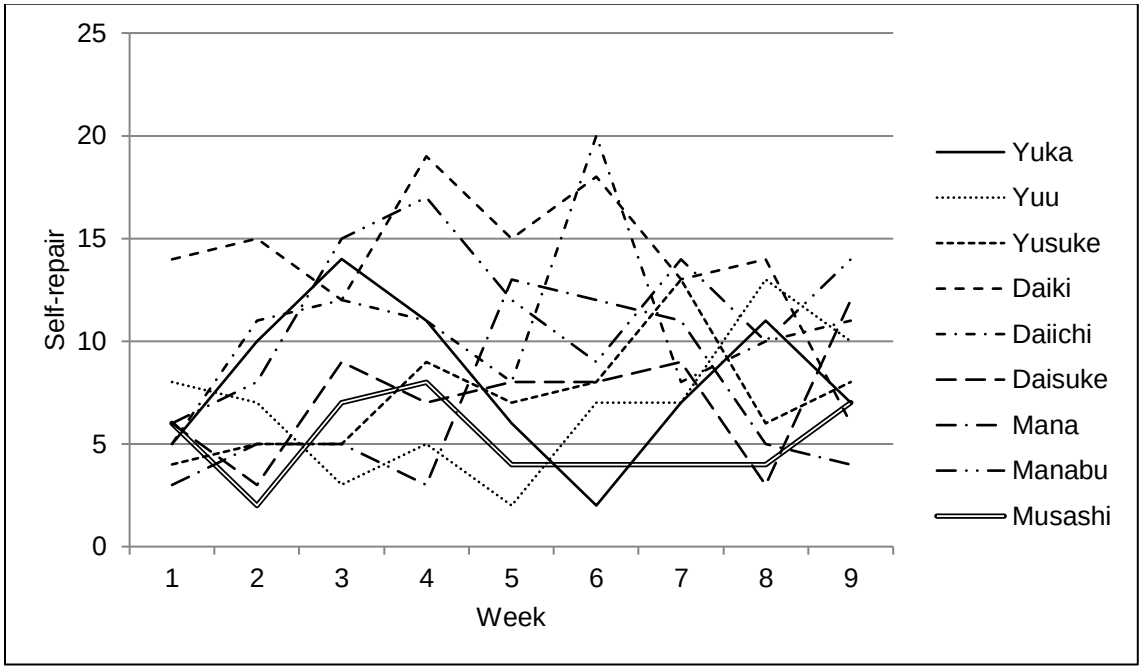



*Figure 16*. Weekly changes in self-repair. Solid lines are Comparison group representatives; Dashed lines are FFI group representatives; Dotted lines are FFI = PI group representatives.

Table 57 shows the descriptive statistics of changes in self-repair of the nine participants from Week 1 to Week 9. Six participants increased self-repair in Week 2 compared to Week 1: Yuka (comparison group) increased five, Yusuke (comparison group) increased one, Daiki (FFI group) increased one, Daiichi (FFI group) increased six, Mana (FFI + PF group) increased two, and Manabu (FFI + PF group) increased by two self-repairs. In Week 3, five participants increased in self-repair: Yuka (comparison group) increased four, Daiichi (FFI group) increased two, Daisuke (FFI group) increased six, Manabu (FFI + PF group) increased seven, and Musashi (FFI + PF group)

increased five self-repairs. Many participants made large fluctuations during the weeks, for example, Daiichi increased from eight self-repairs in Week 5 to 20 self-repairs in Week 6 before he decreased to eight self-repairs in Week 7.

Table 57. *Weekly Performance of Self-Repair*

| | Comparison group | | | FFI group | | | FFI + PF group | | |
|---|---|---|---|---|---|---|---|---|---|
| Week | Yuka | Yuu | Yusuke | Daiki | Daiichi | Daisuke | Mana | Manabu | Musashi |
| 1 | 5 | 8 | 4 | 14 | 5 | 6 | 3 | 6 | 6 |
| 2 | 10 | 7 | 5 | 15 | 11 | 3 | 5 | 8 | 2 |
| 3 | 14 | 3 | 5 | 12 | 12 | 9 | 5 | 15 | 7 |
| 4 | 11 | 5 | 9 | 19 | 11 | 7 | 3 | 17 | 8 |
| 5 | 6 | 2 | 7 | 15 | 8 | 8 | 13 | 12 | 4 |
| 6 | 2 | 7 | 8 | 18 | 20 | 8 | 12 | 9 | 4 |
| 7 | 7 | 7 | 13 | 13 | 8 | 9 | 11 | 14 | 4 |
| 8 | 11 | 13 | 6 | 14 | 10 | 3 | 5 | 10 | 4 |
| 9 | 7 | 10 | 8 | 6 | 11 | 12 | 4 | 14 | 7 |

*Note*. *Note*. FFI = Form-focused instruction; FFI + PF = Form-focused instruction and peer feedback; Smaller numbers indicate a decrease in self-repair, which indicates higher fluency.

**Qualitative Changes in Individual Participants**

In this section I qualitatively describe changes in the speaking performance of the nine participants over the nine weeks using the following codes. Dependent clauses were placed in brackets [ …] following Foster et al. (2000). Grammatical errors were underlined. Self-repairs, including repetitions, self-corrections, false starts, and reconstructions were placed into {…}. Reconstructed or false started words or phrases were coded with an em dash, (—), at the end of each reconstructed word or phrase.

**Yuka's changes in speaking proficiency.** In Week 1, Yuka narrated the four-picture cartoon roughly half in past tense and half in present tense; thus, there was no consistency in her choice of verb tense. Unlike the treatment group participants, who

improved in the use of the *be* verb early in the treatment period, Yuka continued to make errors in her use of *be*. For example, Yuka said in Week 4, "The room is very small and the room is full with many things," instead of, "The room was very small and the room was full with many things." In Week 4, Yuka narrated the cartoon mostly in present tense with some use of past tense. In Week 5, Yuka increased the percentage of accurate past tense use as she accurately self-corrected some past tense verb forms. For example, she said, "And she {said that} {she talk} talked her frighten things." Yuka self-corrected the inaccurate simple present tense, *talk*, into the accurate simple past form tense, *talked*. She also self-corrected, *said*, to *talked*. However, unlike the other treatment group participants who used *could* (past tense of can), Yuka continued to use *can* in its present tense form, for example in Week 6, Yuka said, "So, he cannot spend time with his family", and, "A few days later, he can't stand that situation."

Yuka used very few dependent clauses, and all of the dependent clauses she used were noun clauses such as, *she said* [*that*…]. In Weeks 7 and 8, Yuka attempted to use adverbial clauses for the first time but they were grammatically inaccurate. Yuka also produced some very short T-units throughout the nine weeks.

Yuka produced five self-repairs in Week 1, and increased to 10 self-repairs in Week 2. In Week 3, Yuka peaked in the number of self-repairs, as she started to do more self-corrections of *be* verb. For example, she said, "They are very trouble," and then corrected herself and said, "They were very trouble." In Week 5, Yuka decreased to six self-repairs and further decreased to two self-repairs in Week 6. In Week 7 and Week 8, she made numerous attempts to self-correct her use of verb tenses, so the

number of self-repairs increased to seven in Week 7 and 11 in Week 8, respectively. For example, Yuka said, "{But she don't} but she was not satisfied with her works and {She is} she {want} wanted to be a dress maker."

**Yuu's changes in speaking proficiency.** Yuu continued making grammatical errors throughout the nine weeks. He had the highest performance of percentage of accurate past tense verb use in Week 1 and this performance decreased as the weeks progressed and then increased again towards the end of the experiment period. Yuu used mostly simple past tense forms, a small number of past continuous forms, and no past perfect forms. Most of Yuu's past continuous forms were inaccurate. In Week 1, Yuu started with 81.45% of accurate past tense verb use and decreased to 44.44% in Week 2. In Week 2, out of the nine simple past forms he used, five were inaccurate. Meanwhile, Yuu started self-correcting his inaccurate past tense verb uses. For example, he said, "{they shows} they watched the notice board." On the other hand, Yuu also self-corrected originally accurate verbs to inaccurate verbs. For example, in Week 2, he said, *"{a woman suggested} a woman suggest."* In Week 3, Yuu increased the percentage of accurate past tense verb use to 61.54% but then decreased to 37.50% in Week 4; the large fluctuations in the percentage of accurate past tense verbs continued. In Week 6, Yuu produced only one accurate past tense verb, *Everyone agreed his idea*. In Week 8, Yuu produced five past continuous forms, but four of them were inaccurate. In addition, he made several errors with simple past tense where he self-corrected an originally accurate verb to an inaccurate verb; for example, *When* {*his*

*mother*} *his mother bring— brong juices and cookies for them, she asked them to play outside because it is sunny that day*.

Yuu produced more complex and longer T-units than the other eight participants. In Weeks 1 and 2, he used longer, and more complex sentence structures. In Week 2, he used many dependent noun clauses, and in Week 3, he produced some long T-units such as, *But then a moment later, she used the cellphone [her son gave her as a birthday present] and call to emergency serv*ices. Yuu also used a variety of dependent clauses, including adjective clauses and adverb clauses. In Week 3, he produced a complex T-unit with double built-in dependent clauses, *They thought that* [*they should move to a larger place* [*because they have trouble living there*.]]

Yuu produced a smaller number of repetitions and self-repairs throughout the nine weeks when compared to other participants. He also gradually increased his speech rate as the weeks progressed; however, he made a large decrease in Week 7 before increasing his speech rate in Weeks 8 and 9.

**Yusuke's changes in speaking proficiency.** Overall, Yusuke used more action verbs in his narrations compared to the other participants while using less complex, shorter sentences. Among the action verbs he used throughout the nine weeks, he used mostly simple past tense forms, some inaccurate past continuous forms (from Week 7 onwards) and only one past perfect form. Yusuke's percentage of accurate past tense verb use decreased when he increased in the number of self-repairs. Yusuke had the highest percentage of accurate past tense verb use in Week 1 and then his percentage

rapidly decreased in Week 2. In Week 3, he started correcting his own inaccurate past tense verb forms, so his percentage rapidly increased. For example, Yusuke said, “The next day, the old lady {says} said it is too complicated to use.” Yusuke not only corrected his verb forms, but also corrected some vocabulary choices. For example, in the same week, Yusuke said, “Two weeks later, when {old woman} old lady was going go hiking with her friend and he self-corrected, *old woman* to *old lady*. In Week 7, Yusuke had his second highest percentage of accurate past tense verb use (66.67%), second to his pretest (75.00%). In addition, he started using the past continuous tense, but inaccurately.

Yusuke produced many short T-units throughout the nine weeks while also producing complex sentences with dependent clauses. Among the dependent clauses used by Yusuke, almost all clauses used were noun clauses, for example, *she said th*at and *he said that*. Throughout the nine weeks, Yusuke used one adverbial clause.

While Yusuke used comparatively fewer self-repairs than the other participants in the initial weeks, the number of self-corrections increased as the weeks progressed. Yusuke’s number of self-repairs peaked in Week 7 where he produced 13 self-repairs. Within the 13 self-repairs, of which nine were tense related corrections.

**Daiki’s changes in speaking proficiency.** Overall, Daiki improved in all six measurements over time with week-to-week functions. For accuracy, Daiki started with 70.00% accurate use of past tense verbs in Week 1, decreased to 50.00% in Week 2, remained at 50.00% in Week 3, and then steadily increased in the following weeks until

reaching 100.00% in Week 9. In Week 1, Daiki used both present and past tenses in one sentence. For example, Daiki said, “She signed up and she join.” In Week 4, Daiki narrated mostly in the past tense, but he switched back to the present tense when moving from one picture to the next picture in the four-picture cartoon. For example, Daiki said in Picture 1, “So {so} they eee thought that they should move to a {large} large place” and then he said in Picture 2, “And one year later, eee they {move} move to a {very} {very great eee} very new building. {The room was very new} the room was new and beautiful.” In Week 5, Daiki increased the number of his self-corrections when producing different verb forms. For example, he said, “{She was followed} {she was being followed} she was followed.” In Week 6, Daiki narrated mostly in the past tense but switched back to the present tense when moving from one picture to the next picture as he had done in previous Week 5. For example, Daiki said in Picture 1, “{He} he {should have} eee {over} {overtime working} should have had overtime working” and then he said in Picture 2, “Then a few days later, {he} {he confess} he confess his manager {that} eee that he {have} has no time to spend with his family. And {asked} asked him.” In Weeks 5, 6, 7, and 8, Daiki continued to make a few errors with verbs. During these weeks, changes in Daiki’s percentage of error-free T-units were small, but his percentage of accurate use continued to increase steadily between Weeks 7 and 8, and between Weeks 8 and 9.

Daiki used more dependent clauses than the other participants, and he used dependent clauses every week except in Week 7. In Week 2, Daiki produced five dependent clauses, and in Week 3, he produced eight dependent clauses. In addition, in

Week 3 Daiki produced a complex sentence with two dependent clauses, *Two weeks later when* [*she went hiking with her friend*] *and when* [*she was enjoying the hiking*], *she fell down and hurt her leg*.

Daiki's speech rate increased but fluctuated slightly throughout the experiment. In Week 4, the number self-repairs increased from 12 to 18 when he repeatedly self-corrected his narration from present to past tense, as indicated above when he increased the percentage of error-free T-units and percentage of accurate past tense use. Daiki had more repetitions than the other eight participants every week from Week 1 to Week 8. From Week 8 to Week 9, Daiki had a large decrease in the number of repetitions from 14 to six self-repairs. Daiki's number of repetitions decreased while the number of self-corrections gradually increased in the latter half of the research period.

**Daiichi's changes in speaking proficiency.** Overall, Daiichi had extremely low grammatical accuracy but he improved a great deal over time. Daiichi produced a comparatively low percentage of error-free T-units (23.08%) and instances of accurate past tense verb usage (11.11%) in Week 1 compared to the other participants, and he fluctuated greatly throughout the nine weeks. In addition, Daiichi narrated almost the entire cartoon in present tense except one use of simple past tense, {*They are dee*} *they died*. In Week 2, Daiichi also narrated almost entirely in present tense. From Week 3, he started switching between present and past tense; however, his past tense use increased and use of present tense decreased. For example, *Two weeks later, she went hiking* before he began using present tense again, *The mountain is very hard so she slip by*

*stone*. In Week 4, Daiichi narrated the story entirely in the present tense, except in the middle of the narration where he said, “So one year later, they {find} found a very good place.” In Week 5, Daiichi used more past tense forms, but in Week 6, he switched to mostly using present tense. In Week 9, Daiichi still produced a large number of past tense errors, for example he said, “The TV program say” and “So she want to save the animals.” Meanwhile, Daiichi improved noticeably in his use of *be* verbs. For example, he said in Week 8, “It was sunny” and in Week 9, “She was worried.”

Daiichi had the highest number of self-corrections among the nine participants in almost all weeks. In Week 3, he said, “so she {slipped at} slip by stone”, while the word *slipped* was used correctly, Daiichi corrected and replaced *at* with *by*, and did not include *ed* with the second *slip*. In Week 6, Daiichi made 14 self-corrections, which was his highest number of self-corrections and also the highest number among all participants. Daiichi also had large increases and decreases in speech rate throughout the nine weeks. There was no trend in complexity in Daiichi’s narrations except he did not use dependent clauses.

**Daisuke’s changes in speaking proficiency.** In Week 1, 44.44% of Daisuke’s T-units were error-free and this percentage steadily increased over the following weeks until it reached a peak of 92.86% in Week 6. It then slightly decreased to 70.00% in Week 7, increased to 85.71% in Week 8, and decreased to 63.64% in Week 9. In addition, Daisuke produced accurate past tense forms 50.00% of the time in Week 1 and this figure decreased to 33.33% in Week 2, increased to 75.00% in Week 3, and his

performance fluctuated slightly until it reached 100.00% in Week 7 and stabilized. In Weeks 1 and 2, Daisuke narrated the pictures mostly in the present tense. In Week 3, Daisuke started narrating pictures in the past tense and used *was able to* and *could* instead of *can*. Daisuke did not produce as many dependent clauses as the other participants and the variation in the types of dependent clauses produced was also limited. Daisuke mostly produced noun clauses such as, *she said that*. The only different dependent clause appeared in Week 6, when he said "A month later, [because of this, he was able to come home early and he could enjoy family time." Daisuke was one of only two participants who decreased in speech rate in the posttest (57.36 spm) compared to the pretest (59.17 spm). Daisuke's speech rate increased to 72.47 spm in Week 2 and then stabilized over the following weeks with slight fluctuations before it decreased to 57.29 spm in Week 7, increased to 72.14 spm in Week 8, and then finally decreased to 57.36 spm in Week 9. Daisuke's self-repairs also increased from six on the pretest to 12 on the posttest. Among the self-repairs, Daisuke had the highest number of self-corrections in Week 6 when he corrected his use of verb tenses. For example, he said, "{He} his children {had slept} had already slept."

**Mana's changes in speaking proficiency.** In Week 1, Mana used a mixture of present and past tenses and produced accurate past tense verb forms 62.50% of the time. Mana's percentage of accurate past tense verb use fluctuated greatly where it decreased to 16.67% in Week 2 and then increased to 75.00% in Week 3. From Weeks 4 to Week 9, Mana's percentage of accurate past tense verb use stabilized and ranged mostly

between 80.00% and 90.00%. Mana also produced a noticeable number of plural and third person –*s* errors in the initial weeks. Mana produced error-free T-units 30.00% of the time in Week 1 and this percentage increased steadily over the following weeks until it reached 81.82% in Week 4 and then stabilized. In Week 5, Mana made a noticeable change when she started using *could* instead of *can*. This change was evident in Week 5 where Mana said, "A month later, the citizen patrol {was so succeed} was successful and {a woman a woman went a woman can go home with safe safety. A woman was go home} a woman could go home safely." Mana retained her use of *could* in Week 6 where she said, "Thanks to that decision, a month later, he could go home earlier at about seven o'clock. And {he could eat} he could eat dinner {with} with his family and talk. And finally {he could} he get much time to spend with his family." However, in Week 8, Mana reverted back to using *can* when she said, "They cannot play soccer game in the park."

Mana did not initially use complex T-units; she started producing more complex sentences in the later weeks. From Weeks 6, 7, and 8, Mana used at least one complex sentence structure per week involving one or more types of dependent clauses. In Week 6, Mana used an adverb clause when she said, "So it is important to ask some problem, not only in yourself but people around you." In Week 8, Mana produced the past continuous tense form six times, which was the highest number of past continuous verb tense forms used in a single narration among all the participants in the study.

Mana made no self-corrections in Week 1 and then started making self-corrections in Week 2. The number of self-repairs continued to increase until it reached

its highest number in Week 5 where Mana started using *could*. Mana continued to have a large number of self-repairs in Weeks 6 and Week 7 before the number of self-repairs noticeably decreased in Weeks 8 and 9 where she only made one self-correction in each week. Mana's speech rate fluctuated strongly in the initial weeks. Her speech rate was the highest in Week 1, it decreased by 38.26% in Week 2 and increased by 55.01% in Week 3. From Week 4, Mana's speech rate stabilized; her fluctuations remained 10.00% or less for the remaining weeks. Mana decreased the most in speech rate over the nine weeks among the participants. She decreased 12.13% on the posttest compared to the pretest; however, she made the largest percentage increase in the total number of syllables over the nine weeks among the nine participants.

**Manabu's changes in speaking proficiency.** In Week 1, Manabu used a mixture of present and past verb tenses. In Week 2, he still used both present and past tenses, but the frequency at which he produced present tense decreased. From Week 3, most of the tenses he produced were accurate when Manabu self-corrected himself from the present to the past tense. For example, Manabu said, "{and then she explain} then he explained." However, there were incidents in which Manabu self-corrected himself from accurate to inaccurate tenses. For example, Manabu said "{She said} she complain her son, that it is too complicated to use. I don't need a cellphone." While the first verb *said* was accurate, Manabu inaccurately replaced it with *complain*. In Weeks 5 and 6, Manabu used past perfect tense for the first time, but both uses were incorrect. In Week 7, Manabu produced two past perfect forms and he was accurate in both instances. In

the initial weeks, Manabu's errors were related more to tense errors, while in the later weeks, Manabu's errors happened when he used more complex grammatical structures in longer sentences.

Manabu had the highest speech rate among the nine participants; his Week 1's speech rate was roughly twice the speech rate of Daisuke's (FFI group). In Week 2, Manabu slightly decrease his speech rate and in Week 3, his speech rate further decreased. In Week 4, Manabu's speech rate stabilized and remained stable for the remaining weeks until he made a slight increase in Week 9. Although Manabu's speech rate decrease on the posttest compared to the pretest, the total number of syllables he produced increased on the posttest. Manabu had a large number of self-repairs, between Weeks 3 to 5 and Weeks 7 to 9. In the initial weeks, Manabu self-corrected himself from present to past tense, while in the later weeks, he self-corrected himself while adding more content into the storyline or replacing high-frequency vocabulary and sentence structures with lower-frequency vocabulary and structures. In Week 8, Manabu said, "{because they interrupted in} because they interrupted for someone" where he self-corrected *in* to *for*.

Manabu used more complex sentence structures than the other participants, of which most were accurate. In Week 2, Manabu narrated complex sentences that included multiple dependent clauses in one T-unit. For example, he said, "And a few weeks later, the mayor acted the policy which, [in the city hall to reduce energy cost, [we should turn off all light during lunchtime.]]" Manabu also used more past continuous and past perfect verb forms than the other participants. In Week 3, he

continued to use more complex structures. For example, *And because her friend couldn't use the cellphone.* [*A moment later, she could the cellphone*] and *Because of his son explanation of using the cellphone.* [*She could manage to solve the trouble*]. In Week 5, Manabu said, "all citizens could walk around the city in the dark night even in the dark night safely." Similar to previous weeks, Manabu self-corrected himself in order to add more content to his narration, where he said "all citizens could walk around the city in the dark night" and then self-corrected himself by saying "even in the dark night safely" to emphasize the tone of his narration. Manabu strongly outperformed the other participants on mean length of T-unit in the initial and final weeks. For example, In Week 1, Manabu performed mean length of T-unit of 13.00 and in Week 2, it increased to 15.00. In Week 8, he produced a mean length of T-unit of 13.82 and it was 13.47 in Week 9.

**Musashi's changes in speaking proficiency.** In Week 1, Musashi accurately used past tense 63.64% of the time. He narrated mostly in the past tense with some present tense forms. In Week 2, Musashi decreased to 50.00%, as roughly half of the story was in the past tense. In Week 3, Musashi was 100.00% accurate with past tense because he began to self-correct his errors. For example, Musashi said, "her friend slipped and {hurt} {hurted} hurt her knees." Musashi sustained his 100.00% accuracy with the past tense in Weeks 5 and 6. In week 7, Musashi's percentage of accurate past tense verb usage decreased when he narrated in present tense, "she work at desk, but she have what she want to do." In Week 8, his percentage of accurate past tense verb usage

increased. For example, he said, “they are were prohibited from playing the soccer.” However, Musashi continued to produce errors with past continuous tense usage when he said, “So, a little later, the man who driving car shouted, [move out of the way.] So, the next day, when the boy’s mother walking in front of the park, [she watched his son and his friends.]” In addition, in Week 8, Musashi used both the present and past tenses in one T-unit structure, *they sit chair and played game*.

Musashi displayed almost no change in the number of self-repairs. He made six self-repairs on pretest and seven on posttest. Weekly changes were also small. Musashi also had the smallest number of repetitions among the nine participants. He produced only one repetition each in Weeks 1 and 2. In Week 3, the number of self-corrections increased to four self-repairs. Moreover, there was also almost no change for speech rate from the pretest to the posttest. His speech rate was 76.46 spm in Week 1, 73.71spm in Week 2, it decreased to 59.38 spm in Week 3, and gradually increased again in Weeks 4 and 5. His speech rate decreased to 54.87 spm in Week 7 and gradually increased to 65.38 spm in Week 8 and 74.68 spm in Week 9.

Musashi used complex structures and sentences with dependent clauses in Week 1. In Week 2, Musashi produced numerous short T-units that started with *so*. For example, he said, “|so they use too many electricity |so electricity cost is very high. |It is too costly. |So the mayor decide to reduce the electricity cost. |So during lunch hour, all lights should off. |So, the next day at lunchtime, there is no employee at city hall and nobody use the electricity.|” In Week 2, Musashi used noun dependent clauses on two occasions. In Week 3, Musashi continued to use short T-units. He used two noun

dependent clauses, *she said* [*I don't need a cellphone.*] *Her son felt sad. But he said,* [*you should have a cellphone in case you are in danger*.] In Week 4, Musashi produced the largest number of past perfect forms in a single narration among the nine participants and he used the tense accurately. In Week 5, Musashi used the past perfect tense accurately. In addition, he produced five dependent clauses. In Week 6, Musashi narrated, "{he could enjoy lunchtime lunchtime} {janai} he could enjoy dinner with his family." In Week 7, Musashi said "she work at desk, but she have what she want to do" where he switched from past tense back to present tense. Musashi rarely used past continuous tense throughout the nine weeks.

### Research Hypothesis 4: Differences in Learner Affect

Hypothesis 4a stated that the FFI + PF group would increase in classroom English speaking anxiety more than the FFI and comparison groups, and the FFI group would increase more than the comparison group because FFI + PF group received more form-focused instruction. Hypothesis 4b stated that the FFI + PF group would increase speaking self-efficacy more than the FFI and comparison groups, and the FFI group would increase more than the comparison group because form-focused instruction could increase self-confidence (Van Batenburg et al., 2019). Hypothesis 4c stated that the FFI + PF group would decrease in desire to speak English more than the FFI and comparison groups, and the FFI group will decrease more than the comparison group because form-focused instruction is likely to decrease learners' willingness to communicate (Van Batenburg et al., 2019).

In order to examine Hypothesis 4, three one-way ANCOVAs were conducted for the three affective variables respectively (classroom English speaking anxiety, English speaking self-efficacy, and desire to speak English). The independent variable was group (three levels: comparison, FFI, and FFI + PF). Three ANCOVAs were chosen for this research hypothesis because the Pearson correlation analysis among the three affective variables at the pretest indicated that the variables were moderately correlated (Table 58). Classroom English speaking anxiety and English speaking self-efficacy correlated at $r = -.53$, $p < .001$, classroom English speaking anxiety and desire to speak English correlated at $r = -.27$, $p < .001$, desire to speak English and English speaking self-efficacy correlated at $r = .38$, $p < .001$. Previous research suggested that MANOVA function most effectively when the correlation among dependent variables are highly negatively correlated (Field, 2009; Tabachnick & Fidell, 2007).

Table 58. *Correlations Between Three Affective Variables*

| | 1 | 2 | 3 |
|---|---|---|---|
| 1. Classroom English speaking anxiety | — | | |
| 2. English speaking self-efficacy | -.53* | — | |
| 3. Desire to speak English | -.27** | .38** | — |

** $p < .001$; *$p < .01$

Before conducing the ANCOVA, the normality of the distribution was checked by converting Rasch person ability estimates into *z*-scores. Values above ± 3.29 were deleted. Second, the homogeneity of variance assumption was checked and met using Levene's test. Lastly, the homogeneity-of-slopes assumption was checked and met for the one-way ANCOVAs. Table 59 presents the descriptive statistics of the gain in

classroom English speaking anxiety for the three groups. There were no univariate outliers at $z > \pm 3.29$ (Field, 2009). A normal distribution was assumed for the classroom English speaking anxiety measure because both skewness and kurtosis statistics were under the ±2.00 criterion (George & Mallery, 2016). Both the FFI and the comparison groups decreased in classroom English speaking anxiety while the FFI + PF group increased slightly.

Table 59. *Descriptive Statistics of Classroom English Speaking Anxiety Gain*

| | Comparison group | FFI group | FFI + PF group |
|---|---|---|---|
| *M* | -0.43 | -0.56 | 0.07 |
| *SE* | 0.19 | 0.15 | 0.15 |
| 95%CI | [-0.80, -0.06] | [-0.89, 0.25] | [-0.23, 0.37] |
| *SD* | 0.89 | -0.63 | 0.81 |
| Skewness | -0.39 | -0.13 | -0.02 |
| SES | 0.46 | 0.38 | 0.38 |
| Kurtosis | -0.04 | 1.22 | 1.52 |
| SEK | 0.90 | 0.75 | 0.75 |

*Note*. *Note*. FFI = Form-focused instruction; FFI + PF = Form-focused instruction and peer feedback; All statistics are based on Rasch logits; CI = 95% confidence interval; SK = skewness; KT = kurtosis; SES = Standard error skewness; SEK = Standard error kurtosis.

Table 60 presents the descriptive statistics of the gains (posttest - pretest) in English speaking self-efficacy for the three groups. Two univariate outliers with *z*-scores > ± 3.29 were deleted (Field, 2009). A normal distribution was assumed for English speaking self-efficacy, as the skewness and kurtosis statistics of all three groups were under ±2.0 (George & Mallery, 2016). All three groups increased in English speaking self-efficacy; the comparison group increased the most, followed by the FFI group, and the FFI + PF group.

Table 60. *Descriptive Statistics of English Speaking Self-Efficacy Gain*

| | Comparison group | FFI group | FFI + PF group |
|---|---|---|---|
| *M* | 1.23 | 0.44 | 0.93 |
| *SE* | 0.36 | 0.21 | 0.29 |
| 95%CI | [0.53, 1.94] | [0.01, 0.89] | [0.44, 1.43] |
| *SD* | 2.16 | 1.33 | 1.51 |
| Skewness | 1.23 | 0.26 | 0.25 |
| SES | 0.46 | 0.37 | 0.38 |
| Kurtosis | 1.88 | 0.81 | 0.68 |
| SEK | 0.90 | 0.73 | 0.75 |

*Note*. *Note*. FFI = Form-focused instruction; FFI + PF = Form-focused instruction and peer feedback; All statistics are based on Rasch logits; CI = 95% confidence interval; SK = skewness; KT = kurtosis; SES = Standard error skewness; SEK = Standard error kurtosis.

Table 61 presents the descriptive statistics of the gain (posttest - pretest) in desire to speak English for the three groups. One univariate outlier with a *z*-score of > ± 3.29 was deleted (Field, 2009). Despite deleting the univariate outlier, a normal distribution was not assumed for desire to speak English, as the kurtosis statistic for the FFI + PF group was over 2.00 (George & Mallery, 2016). The comparison group decreased in desire to speak English, the FFI group made almost no change, and the FFI + PF group increased slightly.

Table 61. *Descriptive Statistics of Desire to Speak English Gain*

| | Comparison group | FFI group | FFI + PF group |
|---|---|---|---|
| *M* | -0.92 | -.010 | 0.26 |
| *SE* | 0.42 | 0.33 | 0.28 |
| 95%CI | [-1.75, -0.09] | [-0.74, 0.55] | [-0.23, 1.11] |
| *SD* | 2.00 | -0.74 | 1.70 |
| Skewness | -0.05 | 0.55 | 0.49 |
| SES | 0.46 | 2.18 | 0.39 |
| Kurtosis | 0.89 | -0.18 | 3.39 |
| SEK | 0.90 | 0.37 | 0.76 |

*Note*. *Note*. FFI = Form-focused instruction; FFI + PF = Form-focused instruction and peer feedback; All statistics are based on Rasch logits; CI = 95% confidence interval; SK = skewness; KT = kurtosis; SES = Standard error skewness; SEK = Standard error kurtosis.

Before conducting the ANCOVAs, three one-way ANOVAs were conducted on the pretest of the three affective variables to ensure that there were no significant differences between the three groups. The independent variable was group (three levels: comparison, FFI, and FFI + PF) and the dependent variable was the classroom English speaking anxiety posttest represented by Rasch person ability estimates. The one-way ANOVA was not significant, $F = (2, 96) = 1.13$, $p = .33$. Therefore, it was established that the groups were not significantly different at pretest. The above analysis was repeated for desire to speak English and English speaking self-efficacy. The one-way ANOVA for desire to speak English was not significant, $F(2, 96) = 1.92$, $p = .15$, nor was the one-way ANOVA for English speaking self-efficacy, $F(2, 96) = .96$, $p = .39$, which confirmed that the groups were not significantly different in their English speaking self-efficacy and desire to speak English on the pretest.

The first ANCOVA was conducted to investigate the effects of form-focused instruction by setting group as the independent variable (three levels: comparison, FFI, and FFI + PF) and the classroom English speaking anxiety posttest as dependent variable using Rasch person ability estimates. The covariate was the classroom English speaking anxiety pretest as expressed in Rasch person ability estimates. The ANCOVA was significant for classroom English speaking anxiety, $F(2, 95) = 5.67$, $p < .05$, $\eta^2 = .11$. Follow-up post hoc tests were conducted to evaluate pairwise differences among the three groups. Because the sample sizes differed, Hochberg's GT2 post hoc test was used. The FFI + PF group significantly increased in classroom English speaking anxiety more than the FFI group ($p = .03$) and more than the comparison group ($p = .002$). The

FFI group decreased in classroom speaking anxiety more than the comparison group, but it was not significant ($p = .24$).

The second ANCOVA was conducted to investigate the effects of form-focused instruction by setting group as the independent variable (three levels: comparison, FFI, and FFI + PF) and the English speaking self-efficacy posttest as expressed by Rasch person ability estimates as the dependent variable. The covariate was the English speaking self-efficacy pretest as expressed by Rasch person ability estimates. The ANCOVA was not significant for English speaking self-efficacy, $F(2, 95) = 1.92$, $p = .15$, $\eta^2 = .04$.

The third ANCOVA was conducted to investigate the effects of form-focused instruction by setting group as the independent variable (three levels: comparison, FFI, and FFI + PF) and the desire to speak English posttest as expressed by Rasch person ability estimates as the dependent variable. The covariate was the desire to speak English pretest as expressed by Rasch person ability estimates. The ANCOVA was significant for desire to speak English, $F(2, 95) = 10.78$, $p < .01$, $\eta^2 = .18$. Follow-up post hoc tests were conducted to evaluate pairwise differences among the three groups. Because the sample sizes differed, Hochberg's GT2 post hoc test was used. The FFI + PF group increased in desire to speak English significantly more than the comparison group ($p = .001$) and also increased more than the FFI group, but the difference was not significant ($p = .26$). The FFI group increased in desire to speak English significantly more than the comparison group ($p = .001$).

## Research Hypothesis 5: Learner Self-Assessment

Hypothesis 5a stated that the FFI + PF group would have higher self-assessment of syntactic accuracy than the FFI and comparison groups. In addition, the FFI group would have higher self-assessment of syntactic accuracy than the comparison group because more form-focused instruction leads to greater syntactic accuracy gains (DeKeyser, 1998; Goh & Burns, 2012; Norris & Ortega, 2000; Spada & Tomita, 2010) and therefore results in higher self-assessment (Ma & Winke, 2019; Malabonga et al., 2005; Poehner, 2012). In addition, Hypothesis 5b stated that the interventions would not cause higher self-assessments in terms of syntactic complexity nor oral fluency compared to the comparison group because the form-focused instruction in this study does not target improvements in syntactic complexity nor oral fluency.

In order to examine Hypothesis 5, a MANOVA followed by three one-way ANOVAs were conducted. The independent variable was group (three levels: comparison, FFI, and FFI + PF) and the dependent variables were self-assessed syntactic complexity, self-assessed syntactic accuracy, and self-assessed oral fluency. The MANOVA and ANOVA analyses were chosen because the Pearson correlation analysis of the three self-assessed constructs at pretest indicated that all three variables correlated moderately with each other (Table 62). Self-assessed syntactic complexity and self-assessed syntactic accuracy correlated at $r = .45$, $p < .01$, self-assessed syntactic accuracy and self-assessed oral fluency correlated at $r = .48$, $p < .001$, and self-assessed syntactic complexity and self-assessed oral fluency correlated at $r = .54$, $p$

< .001. MANOVA functions effectively when the variables are not highly correlated or uncorrelated as they would not violate multicollinearity (Tabachnick & Fidell, 2007).

Table 62. *Correlations Between Learner Self-Assessed Variables*

| | 1 | 2 | 3 |
|---|---|---|---|
| 1. Self-assessed syntactic complexity | — | | |
| 2. Self-assessed syntactic accuracy | .45* | — | |
| 3. Self-assessed oral fluency | .48** | .54** | — |

** $p < .001$; *$p < .01$

Table 63 shows the differences between the three groups in terms of gain in self-assessed syntactic complexity. There were no outliers with a *z*-score above ±3.29 (Field, 2009) and the normality of the data was assumed, as the kurtosis and skewness statistics were below the benchmark of < ±2.00 (George & Mallery, 2016). Table 63 indicates that all three groups increased in self-assessed syntactic complexity; the FFI group increased the most, followed by the FFI + PF group and the comparison group.

Table 63. *Descriptive Statistics of Learner Self-Assessed Syntactic Complexity Gain*

| | Comparison group | FFI group | FFI + PF group |
|---|---|---|---|
| *M* | 1.28 | 1.74 | 1.37 |
| *SE* | 0.51 | 0.30 | 0.30 |
| 95%CI | [0.47, 2.10] | [1.11, 2.38] | [0.71, 2.03] |
| *SD* | 2.54 | 1.91 | 1.83 |
| Skewness | 0.42 | 0.60 | -0.09 |
| SES | 0.46 | 0.37 | 0.38 |
| Kurtosis | -0.29 | 0.46 | -0.59 |
| SEK | 0.90 | 0.72 | 0.75 |

*Note*. FFI = Form-focused instruction; FFI + PF = Form-focused instruction and peer feedback; All statistics are based on Rasch logits; CI = 95% confidence interval; SK = skewness; KT = kurtosis; SES = Standard error skewness; SEK = Standard error kurtosis.

Table 64 shows the differences between the three groups in terms of gain in self-assessed syntactic accuracy. There were no outliers with a *z*-score above ±3.29 (Field, 2009) and the normality of the data was assumed, as the kurtosis and skewness statistics were below the benchmark of < ±2.00 (George & Mallery, 2016). Table 64 indicates that all three groups increased in self-assessed syntactic accuracy. The FFI group also increased the most followed by the FFI + PF group and the comparison group. There was a small difference between FFI group and FFI + PF group.

Table 64. *Descriptive Statistics of Learner Self-Assessed Syntactic Accuracy Gain*

| | Comparison group | FFI group | FFI + PF group |
|---|---|---|---|
| *M* | 0.84 | 1.35 | 1.21 |
| *SE* | 0.39 | 0.27 | 0.28 |
| 95%CI | [0.13, 1.54] | [0.80, 1.91] | [0.64, 1.79] |
| *SD* | 1.94 | 1.73 | 1.72 |
| Skewness | 0.13 | 0.31 | 0.69 |
| SES | 0.46 | 0.37 | 0.38 |
| Kurtosis | -0.99 | 0.45 | 1.07 |
| SEK | 0.90 | 0.72 | 0.75 |

*Note*. FFI = Form-focused instruction; FFI + PF = Form-focused instruction and peer feedback; All statistics are based on Rasch logits; CI = 95% confidence interval; SK = skewness; KT = kurtosis; SES = Standard error skewness; SEK = Standard error kurtosis.

Table 65 shows the differences between the three groups in terms of gain in self-assessed oral fluency. There were no outliers with a *z*-score above ±3.29 (Field, 2009) and the normality of the data was assumed, as the kurtosis and skewness statistics were below the benchmark of < ±2.00 (George & Mallery, 2016).Table 65 indicates that all three groups increased in self-assessed oral fluency, the FFI group increased the most, followed by the comparison group and the FFI + PF group.

Table 65. *Descriptive Statistics of Learner Self-Assessed Oral Fluency Gain*

| | Comparison group | FFI group | FFI + PF group |
|---|---|---|---|
| *M* | 1.29 | 1.79 | 1.13 |
| *SE* | 0.44 | 0.30 | 0.27 |
| 95%CI | [0.18, 3.36] | [2.20, 3.98] | [1.23, 3.08] |
| *SD* | 2.21 | 2.88 | 2.87 |
| Skewness | 0.22 | 0.67 | 0.13 |
| SES | 0.46 | 0.40 | 0.38 |
| Kurtosis | -0.12 | 0.79 | 0.75 |
| SEK | 0.90 | 0.72 | 0.75 |

*Note*. FFI = Form-focused instruction; FFI + PF = Form-focused instruction and peer feedback; All statistics are based on Rasch logits; CI = 95% confidence interval; SK = skewness; KT = kurtosis; SES = Standard error skewness; SEK = Standard error kurtosis.

Before the MANVOA analysis was conducted, three one-way ANOVAs were conducted on the questionnaire pretest data to ensure that the starting differences were not statistically significant. Group (three levels: comparison, FFI, and FFI + PF) was the independent variable and the self-assessed syntactic complexity pretest as expressed by Rasch person ability estimates was the dependent variable. The ANOVA was not significant, $F = (2, 101) = .05$, $p = .95$. The same ANOVA was conducted for self-assessed syntactic accuracy, $F = (2, 101) = .19$ $p = .83$, and self-assessed oral fluency, $F = (2, 101) = .22$, $p = .80$; neither was significant. Therefore, it was established that the groups were not significantly different at the pretest.

A one-way MANOVA was conducted to investigate the effects of form-focused instruction on three dependent variables: self-assessed syntactic complexity, self-assessed syntactic accuracy, and self-assessed oral fluency. The independent variable was group (three levels: comparison, FFI, and FFI + PF). The assumptions of the analysis were checked and met. A significant difference was found among the three groups, Wilk's $\Lambda = .79$, $F(6,188) = 3.86$, $p = < .001$, partial $\eta^2 = .11$.

Next, three one-way ANOVAs were conducted; one for each of the three variables, with the independent variable being the group and the dependent variables being the gain scores for self-assessed syntactic complexity, self-assessed syntactic accuracy, and self-assessed oral fluency represented in Rasch person measure.

The homogeneity of variance assumption of self-assessed syntactic complexity was examined and the result was not significant, $F(2, 101) = 1.43$, $p = .25$. The ANOVA results for self-assessed syntactic complexity indicated that there was no significant group effect, $F(2, 101) = .50$, $p = .61$, ($\eta^2 = .01$. The homogeneity of variance assumption of self-assessed syntactic accuracy was examined and the result was not significant, $F(2, 101) = .89$, $p = .41$. The ANOVA for self-assessed syntactic accuracy indicated that the group effect was not significant, $F(2, 101) = .67$, $p = .52$, $\eta^2 = .02$. Finally, the homogeneity of variance assumption of self-assessed oral fluency was examined and the result was not significant, $F(2, 101) = 1.24$, $p = .29$. The ANOVA for self-assessed oral fluency indicated there was no significant group effect, $F(2, 101) = 1.26$, $p = .29$, $\eta^2 = .02$.

## Summary

In this chapter, I presented the results for the five research hypotheses. For Hypothesis 1, both the form-focused instruction (FFI) group and the form-focused instruction and peer feedback (FFI + PF) group improved syntactic accuracy as they improved in the percentage of error-free T-units and percentage of accurate past tense verb forms after seven weeks of intervention. The comparison group did not improve

syntactic accuracy after seven weeks of intervention. All three groups did not improve syntactic complexity. Finally, all three groups improved in oral fluency as assessed by speech rate.

The results for Hypothesis 2 indicated that the FFI group improved syntactic accuracy significantly more than the comparison group as assessed by the percentage of error-free T-units and percentage of accurate past tense verb usage. However, there was no significant difference between the FFI + PF group and the FFI group, or between the FFI group and the comparison group. Hypothesis 2 also found that there was no difference in speaking proficiency gains between the three groups after seven weeks when analytically rated by raters.

For Hypothesis 3, the nine participants from the three groups (comparison, FFI, and FFI + PF) fluctuated differently on the six measures. For mean length of T-unit and clause/T-unit ratio, there were no group differences week-to-week over time, as all nine participants fluctuated slightly week-to-week. For percentage of error-free T-units, the FFI group and the FFI + PF group participants outperformed the comparison group participants, while there was no generalizable difference between the FFI group and the FFI + PF group. All six FFI and FFI + PF group participants increased rapidly between Week 1 and Week 3, and then continued to increase with smaller fluctuations until they reached a ceiling effect in Week 7. Meanwhile, the three comparison group participants fluctuated week-to-week in the percentage of error-free T-units throughout the nine weeks. For the percentage of accurate past tense verb use, the FFI + PF group participants outperformed the FFI group participants in the initial weeks, and then the

differences between the two groups decreased as the weeks progressed. There were no group differences for speech rate or for self-repair.

For Hypothesis 4, a significant group effect was found in classroom English speaking anxiety and desire to speak English. For classroom English speaking anxiety, the FFI + PF group increased significantly more than the FFI and the comparison groups. For desire to speak English, the FFI + PF group increased significantly more the comparison group and the FFI group also significantly increased more than the comparison group. There were no significant group differences for English speaking self-efficacy. However, as shown by the descriptive statistics, the English speaking self-efficacy of all three groups increased on the posttest compared to the pretest.

For Hypothesis 5, there were no significant group differences in the gain score of self-assessed syntactic complexity, syntactic accuracy, or oral fluency. Nevertheless, all groups (comparison, FFI, and FFI + PF) increased in their three self-assessments on the posttest compared to the pretest.

In the next chapter I discuss the results for each of these five research hypotheses presented in the present chapter with reference to the existing literature. I also make pedagogical implications based on my interpretation of results.

# CHAPTER 6

# DISCUSSION

In this chapter, I discuss the five research hypotheses for this study by interpreting the results presented in Chapter 5. Finally, I discuss the pedagogical implications and research implications of this study.

## Research Hypothesis 1: Within-Group Speaking Development

Hypothesis 1a stated that the FFI group and the FFI + PF group would improve syntactic accuracy, while the comparison group would not because form-focused instruction explicitly directed the intervention groups to pay attention to form, especially the use of past tense forms (DeKeyser, 1998; Goh & Burns, 2012; Norris & Ortega, 2000; Spada & Tomita, 2010). Hypothesis 1b stated that all three groups would improve syntactic complexity and oral fluency because task repetition has been found to lead to such improvements (e.g., Ahmadian, 2011; Bygate, 2001). In order to examine the hypotheses, within-in group comparisons were conducted for each group by comparing the six CAF measures (See Table 6) from the pretest and the posttest.

The FFI and FFI + PF groups increased the percentage of error-free T-units and percentage of accurate past tense verb usage on the posttest compared to the pretest. For the FFI group, the one-way repeated-measures ANOVA for both measurements were significant at $p < .01$. For the FFI + PF group, the one-way repeated-measures ANOVAs for both measurements were also significant at $p < .01$. The comparison

group improved a little in the percentage of error-free T-units and the percentage of accurate past tense verb usage also improved a little on the posttest compared to the pretest but the repeated-measures ANOVAs for both measurements for the comparison group were not significant.

All three groups improved slightly in mean length of T-unit and clause/T-unit ratio on the posttest compared to pretest. Only the comparison group improved significantly on mean length of T-unit, $p < .01$. For clause/T-unit ratio, the one-way repeated-measures ANOVAs for all three groups were not significant.

All three groups increased speech rate on the posttest compared to the pretest. The FFI group's one-way repeated-measures ANOVA was significant, $p < .01$, as was that of the comparison group, $p < .01$. The FFI + PF group's one-way repeated-measures ANOVA was not significant, $p = .09$. Finally, all three groups increased self-repair on the posttest compared to the pretest; however, these increases were not statistically significant.

First, Hypothesis 1a stated that the intervention groups would improve syntactic accuracy because form-focused instruction leads learners to attend to specific learning objectives by noticing accurate forms (Schmidt, 1990, 2001). The results indicated that the FFI group and FFI + PF group developed greater syntactic accuracy as shown by the one-way repeated-measures ANOVAs for the percentage of error-free T-units and the percentage of accurate past tense verb usage; both the FFI and FFI + PF groups improved significantly, $p < .01$. Form-focused instruction encouraged the participants in both intervention groups to make cognitive comparisons between their erroneous oral

output and the target forms they were instructed to use (Ellis, 1994). According to the Comprehensible Output Hypothesis (Swain, 1985), learners need to comprehend the errors in their own output and be pushed to modify their output to make it more precise and more appropriate. The participants in the two intervention groups were encouraged to produce more accurate past tense forms because of the teachers' grammar instruction and peer feedback. This result supports the findings reported by Sheppard (2006) and Mochizuki and Ortega (2008) that the explicit teaching of linguistic form can lead to greater grammatical accuracy in communicative speaking tasks. On the other hand, the comparison group did not improve grammatical accuracy after seven weeks of intervention in terms of the percentage of error-free T-units and the percentage of accurate past tense verb usage. In addition, the one-way repeated-measures ANOVAs for the percentage of error-free T-units and the percentage of accurate past tense verb usage for the comparison group were not significant. The comparison group participants did not receive form-focused instruction and thus they were not led to notice their grammatical errors (Schmidt, 1990, 2001). Consequently, they did not make their utterances more accurate. The result supports previous studies (e.g., Bygate, 2001) suggesting that repetition does not improve grammatical accuracy. Therefore, Hypothesis 1a was supported.

Second, Hypothesis 1b stated that all three groups would improve syntactic complexity regardless of receiving form-focused instruction or not because the effects of repetition from the weekly narration task would lead to greater syntactic complexity (e.g. Ahmadian, 2011; Bygate, 2001; Gass et al., 1999). Previous researchers looking at

the effects of repetition on speaking proficiency have suggested that repetition leads to improved syntactic complexity because it increases cognitive capacity (Bygate, 2001; Gass et al., 1999). However, in the present study, the FFI group and the FFI + PF group did not significantly increase syntactic complexity after seven weeks according to the mean length of T-units and clause/T-unit ratio. Only the comparison group significantly improved in global complexity, as the one-way repeated-measures ANOVA of the pretest and posttest of mean length of T-unit was significant, $p < .01$. A possible reason was because the participants in FFI and FFI + PF groups were pressured to pay attention to accurate verb tense forms (Bygate, 1998; Ellis, 2005b; Foster & Skehan, 1996). Therefore, they did not increase syntactic complexity due to limited attentional resources already allocated for syntactic accuracy (Skehan, 2009a).

Meanwhile, the comparison group participants significantly improved global complexity possibly because they were not given form-focused instruction and therefore did not allocate limited attentional resources to greater accuracy. Weekly task repetition allowed the comparison group participants to utilize the cognitive work undertaken in their previous task, which freed up cognitive capacity to produce more complex sentence structures (Bygate, 2001). All three groups displayed no significant gains in subordination complexity, as the clause/T-unit ratio for all three groups was not significant. This insignificant result might have occurred because the participants were at an intermediate level of speaking proficiency and clause/T-unit ratio is considered to be a measure of speaking proficiency for advanced-proficiency learners (Norris & Ortega, 2009). Therefore, the first part of Hypothesis 1b was partially supported.

Hypothesis 1b stated that all three groups would improve oral fluency because the weekly repetitive narrative task would lead to greater automatization (Bygate, 1996, 1999, 2001; Bygate & Samuda, 2005). All three groups increased their speech rate on the posttest compared to the pretest. The comparison group and the FFI group increased speed fluency significantly ($p < .01$); however, the FFI + PF group did not significantly increase speed fluency ($p = .09$). Possible reasons for the significant improvement of speech rate for the comparison and FFI groups was that the participants in these two groups became familiar with the narrative genre, the chronological organization of four-picture cartoons, as well as high-frequency lexis and grammatical features used for narrations from completing the narrative task for seven weeks. These elements freed up capacity for retrieving and using new language knowledge, which accelerated language processes (Bygate, 2001) and resulted in more automatized speech (Gass et al., 1999). The FFI + PF group increased in speech rate but was not significant because these participants likely allocated more attention to improving grammatical accuracy, which directed their attention away from fluency (Skehan, 2009a).

The combined effects of grammar instruction on past tense verb forms and peer corrective feedback probably encouraged the FFI + PF group participants to pay more attention to past tense verbs. Importantly, the FFI + PF group participants did not experience a decrease in speed fluency. Therefore, a trade-off effect between syntactic accuracy and oral fluency did not occur even though the participants had limited working memory capacity and probably directed their attention to improving syntactic accuracy as a result of receiving a greater amount of form-focused instruction from

combined grammar instruction and peer corrective feedback (Bygate, 1998; Ellis, 2005b; Foster & Skehan, 1996).

On the other hand, all three groups decreased repair fluency, as shown by increases in the number of self-repairs on the posttest compared to the pretest; however, the ANOVA results for all three groups were not significant. Thus, the second part of Hypothesis 2b was partially supported because the comparison and the FFI groups improved speed fluency.

Form-focused instruction probably did not trigger trade-off effects between syntactic complexity, syntactic accuracy, and oral fluency after seven weeks of intervention. For the FFI group participants, trade-off effects did not occur between syntactic accuracy and oral fluency. Syntactic accuracy and oral fluency improved simultaneously perhaps because the participants received a sufficient amount of input and output practice over seven weeks of repetitive practice and weekly opportunities to produce oral output. However, the FFI group participants did not improve syntactic complexity; while simultaneous improvement happened with oral fluency and syntactic accuracy, syntactic complexity makes greater cognitive demands on learners (Housen & Kuiken, 2009); therefore, it did not improve simultaneously with syntactic accuracy. The FFI + PF group improved syntactic accuracy but not oral fluency or syntactic complexity. Neither simultaneous improvement (Robinson, 2001) nor trade-off effects occurred given that the FFI + PF group did not decrease in oral fluency nor syntactic complexity when their syntactic accuracy increased.

### Research Hypothesis 2: Effects of FFI on Speaking Development

Hypothesis 2 concerned the effects of form-focused instruction on speaking development by comparing changes in syntactic complexity, syntactic accuracy, and oral fluency of the participants who received FFI, and those who received FFI and peer feedback, as well as those who did not receive FFI. Hypothesis 2a stated that the FFI group will improve syntactic accuracy more than the comparison group because and the FFI + PF group will improve syntactic accuracy more than the FFI group and the comparison groups because form-focused instruction explicitly directs learners to pay attention to form, especially the use of past tense forms (DeKeyser, 1998; Goh & Burns, 2012; Norris & Ortega, 2000; Spada & Tomita, 2010). Hypothesis 2b stated that the FFI + PF and FFI groups will not show greater improvements in syntactic complexity nor oral fluency than the comparison group.

Research hypothesis 2 has two parts. First, the results from the one-way MANOVA indicated that there were significant differences between the comparison group and intervention groups who received form-focused instruction ($p = .008$, partial $\eta^2 = .10$). Among the six follow-up ANOVAs, the percentage of error-free T-units ($p < .01$, partial $\eta^2 = .15$), and the percentage of accurate past tense verb usage ($p < .01$, partial $\eta^2 = .20$) were significant. However, the results of the follow-up post hoc tests indicated that there was no significant difference between the FFI group and FFI + PF group. On the other hand, the other four ANOVAs were not significant: MLT ($p = .20$, partial $\eta^2 = .20$), Clause/T-unit ratio ($p = .51$, partial $\eta^2 = .01$), speech rate ($p = .79$, partial $\eta^2 = .01$), and self-repair ($p = .57$, partial $\eta^2 = .01$).

In the second part, the ANOVA for the analytical ratings just missed significance ($p$ = .06). Nevertheless, the descriptive statistics showed a difference between the three groups: the FFI + PF group increased the Rasch fair average ($M$ = .30), the FFI group made almost no gain ($M$ = .01) and the comparison group decreased slightly ($M$ = -.05).

As hypothesized, the participants who received form-focused instruction improved grammatical accuracy more than the participants who did not; both the FFI and FFI + PF groups improved significantly on percentage of error-free T-units and percentage of accurate past tense verb usage after seven weeks, whereas the comparison group made no improvement. The ANOVAs for both percentage of error-free T-units and percentage of accurate past tense verb usage had large effect sizes ($\eta^2$ = .15 and $\eta^2$ = .20, respectively), indicating that form-focused instruction had a strong effect on syntactic accuracy improvement. The interventions led the FFI and FFI + PF groups to attend to syntactic accuracy and particularly to accurate past tense verb forms. This finding supports previous researchers such as Schmidt (1990, 2001), who found that language-focused instruction effectively directs learners' attention to notice specific learning objectives so that they can make cognitive comparisons between their output and the target grammatical forms (Ellis, 1995; Schmidt & Frota, 1986). In addition, it also supports the findings of Mochizuki and Ortega (2008) and Sheppard (2006), who found that explicit teaching leads to greater morphosyntactic accuracy with communicative tasks. Form-focused instruction effectively develops syntactic accuracy because it directs students' attention to target grammatical features in a structured

environment through demonstrations, explanations, and practice. In this study, the participants in the two treatment groups were given instruction on three past tense verb tenses. Each week, they repeated an equivalent four-picture narration task twice and received form-focused instruction between the two narrations. This explicit form-focused instruction allowed the treatment group participants to reflect on what they had said in the first narration using the instruction they were given before narrating the second time.

However, while form-focused instruction effectively led the participants to develop greater syntactic accuracy, the amount of form-focused instruction given did not make a difference. The FFI group participants received 10 minutes of form-focused grammar instruction, while the FFI + PF group participants received 10 minutes of grammar instruction and 10 minutes of corrective peer feedback. However, there were no significant differences between the two groups for percentage of error-free T-units and percentage of accurate past tense verb usage. This finding refuted Hypothesis 2b, as the FFI + PF group was expected to outperform the FFI group because they received a longer intervention.

Three reasons can be suggested why the greater amount of form-focused instruction did not result in greater syntactic accuracy gains. First, the participants in both intervention groups reached a ceiling effect as a small amount of form-focused instruction might have been sufficient for improving past tense verb errors. Although many Japanese learners of English make past tense errors when speaking, they make fewer such errors when writing. More past tense verb errors in speaking can partly be

attributed to the fact that speaking is an on-line task that requires immediate output while writers have more time to access explicit knowledge. The participants in this study were intermediate-proficiency Japanese university students with high academic ability who had studied the past tense in secondary school. Because the participants had studied past tense verbs previously, even a small amount of form-focused instruction might have been sufficient to raise their awareness of accurate past tense forms.

Second, the syntactic accuracy gains were only focused on changes on the simple past tense form as the past continuous and past perfect forms were excluded from the analyses due to their infrequent use. Less than 5% of the verb forms were past continuous and less than 1% were past perfect on the posttest. As suggested by the first reason, a small amount of form-focused instruction might have been sufficient for improving form-related errors, so a longer intervention time was not needed. If the research focus had been on past continuous or past perfect verb tense errors, then a greater amount of form-focused instruction might have generated greater syntactic accuracy gains.

Third, a greater amount of instruction might have been more effective if the research period had been shorter. The positive effects of form-focused instruction on syntactic accuracy development might have been reached before the final week of the research period. This speculation is examined in the discussion of Hypothesis 3 concerning the weekly trajectory changes of speaking proficiency development.

On the other hand, form-focused instruction did not lead to improvements in syntactic complexity and oral fluency, as there were no significant differences between

the three groups in mean length of T-unit, clause/T-unit ratio, speech rate, and self-repairs. This result was expected because form-focused instruction specifically targets syntactic accuracy (Mochizuki & Ortega, 2008) by directing learners to pay attention to producing more accurate grammatical forms. Therefore, the results of this study do not support the trade-off effects between complexity, accuracy, and fluency (Bygate, 1998; Ellis, 2005a; Foster & Skehan, 1996). The Trade-off Hypothesis states that attentional capacity and working memory are limited; therefore, spending more attentional resources on one aspect negatively impacts one or both of the other aspects (Skehan, 2009a). Fluency competes with accuracy, while accuracy competes with complexity (Housen & Kuiken, 2009). Therefore, some previous researchers have argued that when participants improve grammatical accuracy, they compensate and decrease attention to fluency and/or syntactic complexity (Ellis, 2009a). However, in the present study, the comparison group did not outperform the two intervention groups where syntactic complexity and oral fluency was concerned. The results might have differed if more complex grammatical structures (e.g., relative clauses) had been the focus of the study.

Finally, the analyses conducted on the analytical ratings indicated that there was no significant difference among the three groups, which suggests that the form-focused interventions did not produce greater speaking proficiency as perceived by the raters. Several reasons can be suggested for this finding. First, the raters were less sensitive at capturing the effects of grammar instruction on the learners' speaking performances because it was an impressionistic judgment of overall oral performance (Nitta & Nakatsuhara, 2014). Although form-focused instruction affected specific aspects of the

speaking performances, it did not impact overall speaking proficiency. Second, improvements in the use of past tense might have only slightly contributed to the raters' perceptions of the participants' overall speaking performance (Sawaki, 2007). The analytical ratings were influenced by the raters' perceptions of different weightings of the complexity, accuracy, and fluency subcomponents (Ling et al., 2014). Third, the analytical ratings for syntactic accuracy might have been influenced by the ratings for complexity and fluency (Xi & Mollaun, 2006). Analytically ratings placed high cognitive load on raters when they assessed the three subcomponents of complexity, accuracy, and fluency simultaneously so it might have been difficult for raters to rate the components independently from the others (Xi & Mollaun, 2006). Fourth, the analytical ratings might have been significant if non-linguistic subcomponents had been included. It has been argued by researchers (e.g., Sato, 2011) that content contributes significantly to oral performances.

### Research Hypothesis 3: Effects of FFI on Trajectory of Changes

Hypothesis 3 concerned the changes in the speaking performances of the participants based on the effects of form-focused instruction over time. In order to better understand the effects of FFI on speaking proficiency, the changes of the participants who did not receive FFI and changes of participants who received both FFI and peer feedback were also examined. Hypothesis 3 stated that the FFI + PF group would improve syntactic accuracy more quickly week by week than the FFI group and the comparison group because FFI + PF group received the greatest amount of form-

focused instruction, which can lead to increases in syntactic accuracy (Ahmadian & Tavakoli, 2011; Mochizuki & Ortega, 2008; Williams, 2005; Yuan & Ellis, 2003).

Three representative participants from the comparison, FFI, and FFI + PF groups were selected and changes in their weekly speaking performances were examined using the six CAF measurements. In addition, the participants' weekly narration transcriptions were examined qualitatively. The discussion of weekly changes of these nine participants was divided into three sections: syntactic complexity, syntactic accuracy, and oral fluency.

**Discussion of Weekly Changes in Syntactic Complexity**

Overall, there were no noticeable differences between the nine participants over the nine weeks in terms of mean length of T-unit and clause/T-unit ratio. All nine participants fluctuated slightly week-to-week on both variables. The most noticeable changes happened between Weeks 1 and 2, and between Weeks 2 and 3. All nine participants except Mana (FFI + PF group) and Musashi (FFI + PF group) increased mean length of T-unit between Week 1 and Week 2, and all participants except Daiki (FFI group) and Musashi (FFI + PF group) decreased between Week 2 and Week 3. For clause/T-unit ratio, all participants except Yusuke (comparison group), Mana (FFI + PF group), and Musashi (FFI + PF group) increased between Week 1 and Week 2, and five participants decreased between Week 2 and Week 3.

The accuracy-based form-focused instruction had no effect on syntactic complexity, as there were no significant group differences and no noticeable week-to-

week differences in the mean length of T-unit or clause/T-unit ratio of the nine participants. These results differs from those reported by Kanda (2015), who found significant gains in syntactic complexity in low- and intermediate-proficiency learners who repeated the same type of task during one academic year. A possible reason that form-focused instruction had no effect on syntactic complexity over the nine weeks is because the instruction given in this study was targeted on past tense use, so the participants' attention was not directed at producing longer, more complex sentences. Form-focused instruction can encourage learners to attend to forms that they have not previously noticed (Schmidt, 1990, 2001). It can also help learners to make comparisons between their current interlanguage knowledge and the target language rules by processing well-formed input (Doughty, 2001). Therefore, unless learners notice complex structures in the input they were given, they will not produce more complex sentence structures (Doughty, 2001).

Repetition had immediate short-term effects on syntactic complexity development because seven participants improved mean length of T-unit and six participants improved clause/T-unit ratio between Week 1 and Week 2. This finding supports many previous studies that found positive effects of task repetition on syntactic complexity (e.g., Ahmadian, 2011; Bygate, 2001). Task repetition affects cognitive processes in learners so that cognitive work undertaken on previous occasions is still accessible in memory. As the result, repetition frees up cognitive capacity, so learners can focus on other aspects of the task (Bygate, 2001; Gass et al., 1999). The narrative task in Week 1 allowed the participants to become familiar with narrating a four-picture

cartoon, which freed up their cognitive capacity and allowed them to produce longer, more complex sentence structures in the equivalent narrative task in Week 2. However, task repetition had only short-term effects on syntactic complexity development as indicated by the seven participants who decreased mean length of T-unit between Week 2 and Week 3. While syntactic complexity increased between Week 1 and Week 2, the effects of repetition task probably did not further free up cognitive capacity in the participants in Week 3.

Task repetition did not result in more subordination; the clause/T-unit ratio had the smallest week-to-week changes among all nine participants. This result supports previous studies (e.g., Bygate, 1996; Lynch & Mclean, 2000) showing no significant gains in syntactic complexity after massed repetition. However, it is difficult to conclude that task repetition does not develop subordination complexity because as discussed above, clause/T-unit ratio is a measure for advanced learners. Therefore, weekly increases in subordination complexity might have occurred over time, but they were not identified by clause/T-unit ratio measure.

The nine participants' syntactic complexity changed little week-to-week; their performance for mean length of T-unit and clause/T-unit ratio stayed fairly constant with fluctuations of less than 10% during the remaining research period. A qualitative examination of the weekly transcripts revealed small changes in the participants throughout the study. The participants who produced more syntactically complex sentences in the initial weeks also used more complex structures in the later weeks. Moreover, the participants who used simpler and shorter sentence structures in the

initial weeks also used simpler sentence structures in the later weeks. Syntactic complexity might have fluctuated slightly over the weeks due to the effects of task complexity (Sasayama, 2016), which can be determined by the storyline and indirect messages. The complexity of each storytelling task partly depends on the topic, structure, and picture quality (Elder et al., 2002).

Form-focused instruction might have interfered with syntactic complexity development, as half of the participants who received form-focused instruction did not improve syntactic complexity even though they engaged in the same weekly narration task as the comparison group participants. However, this finding cannot be explained using the Trade-off hypothesis (Skehan, 2009a) because the participants who received more form-focused instruction increased syntactic accuracy while maintaining their levels of syntactic complexity. Spending more attention on one aspect (i.e., syntactic accuracy), did not negatively impact one or both of the other aspects (i.e., complexity and fluency). This finding is consistent with the claims of the Limited Attention Capacity Hypothesis, and therefore does not refute prior findings (e.g., Skehan, 2009a).

**Discussion of Weekly Changes in Syntactic Accuracy**

Overall, Daiki, Daiichi, and Daisuke (FFI group), and Mana, Manabu, and Musashi (FFI + PF group) outperformed Yuka, Yuu, and Yusuke (comparison group) in the percentage of error-free T-units and percentage of accurate past tense verb use over the nine weeks. All treatment group participants made large week-to-week increases.

There was no generalizable differences between the three FFI group participants and three FFI + PF group participants for percentage of error-free T-units. Most of the participants increased rapidly between Week 1 and Week 3. From Week 4 to Week 7, the intervention group participants continued their increases but to a smaller extent and with fluctuations. In Week 7, most of them peaked in their performance of percentage of error-free T-units. Meanwhile, all three comparison group participants decreased in Week 2, increased in Week 3, and then fluctuated throughout the remaining weeks.

All treatment group participants increased percentage of accurate past tense verb use throughout the nine weeks with some fluctuations. The FFI + PF group participants outperformed the FFI group participants in the initial weeks, and then the differences between the two groups decreased as the weeks progressed. All intervention group participants achieved high performances by Week 9. By Week 3, the FFI + PF group participants had already performed well. The FFI + PF group participants continued to increase from Week 4 to Week 9, but the increases were trivial. The FFI group participants continued to increase in percentage of accurate past tense verb use with some week-to-week fluctuations until Week 9.

Form-focused instruction had long-term positive effects on developing syntactic accuracy, as all treatment group participants increased the percentage of error-free T-units and the percentage of accurate past tense use over the nine weeks with some fluctuations. This finding supports the results of Hypotheses 1 and 2, and previous studies such as those by Schmidt (1990, 2001), Sheppard (2006), and Mochizuki and

Ortega (2008), who found that language-focused instruction develops syntactic accuracy because it directs learners' attention to specific learning objectives.

The impact of form-focused instruction on global syntactic accuracy development was most effective in the beginning of the nine-week research period, as the intervention group participants improved most rapidly in the percentage of error-free T-units between Week 1 and Week 3. The finding that form-focused instruction was most effective at the beginning of the nine-week research period differed from Kanda (2015), who found longitudinal form-focused instruction to be most effective between Week 4 and Week 8. Form-focused instruction had faster effects in the present study compared to Kanda's study because the participants in the present study were given more explicit grammar instruction and peer feedback explicitly directed the participants' attention to the accurate use of past tense (Doughty & Williams, 1998). Kanda applied a more holistic definition of form-focused instruction and applied repetition as an intervention. Therefore, while Kanda found gains in complexity, fluency, and accuracy, the gains were more significant in the latter half of her study. Repetition can be one type of form-focused instruction at raising learners' awareness for accurate grammatical forms, but it lacks explicitness in guiding learners to notice the errors that they have made and explicitly lead them to notice what they need to improve (Williams, 2005). However, the diminishing effects of form-focused instruction on grammatical accuracy development can also be explained using the diminishing-returns learning curve theory that the rate of increase in learning is higher in the beginning but

decreases with time until it reaches zero when learners have reached the maximum level of learning and cannot improve further.

After Week 3, form-focused instruction continued to produce improvements in global syntactic accuracy as indicated by the upward trend in the percentage of error-free T-units. Therefore, form-focused instruction continued to have positive week-to-week effects on improving global syntactic accuracy even in the latter half of the research period. However, the effectiveness of instruction on improving global syntactic accuracy stopped in Week 7, which is when the treatment group participants peaked in their performances. The positive effect of the interventions until Week 7 is similar to the finding reported by Lambert, Kormos, and Minn (2017), who found that the effectiveness of intervention on improving speaking fluency continued until the fifth repetition. However, Lambert et al.'s study was focused on the development of oral fluency rather than syntactic accuracy development.

While global syntactic accuracy continued to increase until Week 7, there were week-to-week fluctuations in the percentage of error-free T-units throughout the nine weeks. This result can be explained by task effect as the weekly equivalent narration task involved different content and lexis and therefore generated differences in task complexity (Sasayama, 2016) even though the prompts used in this study had been piloted to estimate their difficulty (Obunsha, 2011).

However, while form-focused instruction effectively developed global syntactic accuracy over time, more instruction was not more effective. The differing amounts of instruction given to the participants made no difference in how they changed in week-

to-week general syntactic accuracy as assessed by the percentage of error-free T-units. This result supports the results of Hypothesis 1 that more form-focused instruction was not more effective, as it did not have a more positive impact on global syntactic accuracy development.

Long-term form-focused instruction improved past tense verb use because all treatment group participants improved greatly in the percentage of accurate past tense verb usage. In addition, the three FFI + PF group participants outperformed the three FFI group participants in the percentage of accurate past tense verb usage in Weeks 2 and 3. Furthermore, the FFI + PF group participants reached their maximum performance in Week 7, while the FFI group participants continued to increase in percentage of accurate past tense verb usage until Week 9. The FFI + PF group participants improved faster in past tense verb usage because they received a greater amount of instruction time on past tense verb usage, which drew more of their attention to the target syntactic forms. The greater amount of form-focused instruction allowed the participants more opportunities to notice inappropriate forms in their own language output, so they modified their spoken language and achieved greater accuracy in the targeted grammatical feature. Form-focused instruction encourages learners to make a cognitive comparison between their output and the target forms they were instructed to use (Ellis, 1995). In this study, the participants were given explicit instruction on three past tense forms. Each week, the participants narrated the four-picture cartoon twice and between the repeated narrations, the treatment group participants received interventions. Form-focused instruction gave the participants the opportunity to reflect on what they

narrated during the first performance where I explicitly directed the participants to pay attention on the accurate use of past tense, before they started the second narration. The FFI + PF group participants had more time to reflect on their production as they listened and reviewed other students' narration after they were given explicit grammar instruction. Corrective peer feedback directed the FFI + PF group participants to notice their problematic forms and encouraged them to produce accurate past tense forms (Li, 2010; Lyster & Saito, 2010). The feedback encouraged the participants to notice discrepancies between their own language production and target language forms. Therefore, the FFI + PF group participants improved in the percentage of accurate past tense verb usage faster than FFI group participants.

Although the amount of instruction positively influenced the accuracy of past tense verbs early in the study, the difference diminished as time passed; there was no difference in the percentage of accurate past tense use between the FFI and FFI + PF groups at the end of the seven-week intervention. Therefore, while more form-focused instruction did not produce greater global syntactic accuracy (see also the results of Hypothesis 1) after the seven-week intervention, the results for Hypothesis 3 indicated that more form-focused instruction led to more rapid improvements in the production of past tense than less form-focused instruction; however, the positive effects of more instruction gradually diminished as time passed.

Two out of the three comparison group participants did not improve percentage of error-free T-units or percentage of accurate past tense verb usage during the study because they were not directed to do so. They fluctuated in their week-to-week

percentage of error-free T-units and percentage of accurate past tense verb usage. Similar to the intervention group participants, the comparison group participants were influenced by task effects as the weekly equivalent narration task involved different content and lexis and generated slightly different difficulty levels. Nevertheless, one comparison group participant slightly improved both global and past tense verb usage over the nine-week research period as seen in the small increases in percentage of error-free T-units and percentage of accurate past tense verb use. Thus, repetition alone had small positive effects on syntactic accuracy development (Gass et al., 1999).

**Discussion of Weekly Changes in Oral Fluency**

No group performed better than the other two groups for self-repair over the research period, and all participants displayed large fluctuations. Despite these fluctuations, all nine participants except Yuka (Comparison) and Daiki (FFI) increased the number of self-repairs as time passed. Comparison group participants (Yuka, Yuu, and Yusuke) and two FFI group participants (Daiki and Daiichi) increased speech rate with week-to-week fluctuations until Week 9. On the other hand, the three FFI + PF group participants (Mana, Manabu, and Musashi) had small week-to-week fluctuations but demonstrated almost no changes in speech rate after seven weeks of intervention. In Week 7, almost all participants decreased speech rate.

Form-focused instruction had no effect on week-to-week improvement of speed fluency because while all three comparison group participants gradually improved in speech rate over the nine-week research period with some week-to-week fluctuations,

all three FFI + PF group participants made no improvement. A possible reason was because the comparison group participants did not receive form-focused instruction and therefore were able to direct their limited attentional resources toward fluency. This finding supports the results of Hypothesis 2 and other previous studies (e.g., Ahmadian, 2011; Bygate, 2001; Lynch & Maclean, 2000) that task repetition improves fluency as it leads to more automatized speech (Gass et al., 1999). In addition, the weekly task improved speed fluency because the linguistic knowledge activated in the first task performance was still available when the participants repeated the equivalent task in the following week. Therefore, it freed up the capacity for retrieving and using new forms and speeded up language processing (Bygate, 2001).

Small amounts of form-focused instruction can positively impact speed fluency development, as two of FFI group participants increased in speech rate over the seven-week intervention despite also receiving form-focused instruction. A possible explanation for this finding was that while form-focused instruction might have interfered with the speed at which the FFI group participants were narrating, the positive effects of task repetition were greater and overruled the interference. Therefore, a smaller amount of instruction did not interfere with the week-to-week improvement of speed fluency. The participants in the FFI + PF group did not improve speed fluency because they were directed to pay attention to accuracy. Therefore, they devoted limited attentional resources to fluency development. Nevertheless, while form-focused instruction did not lead to greater speed fluency, it did not have detrimental effects on fluency development given that the FFI + PF group participants changed their speech

rate only slightly over the course of the study. This result for Hypothesis 3 supports the results of Hypotheses 1 and 2 that there was no trade-off effect among complexity, accuracy, and fluency. Notwithstanding limited working memory capacity (Bygate, 1998; Ellis, 2005a; Foster & Skehan, 1996), fluency did not compete with accuracy as suggested by the Trade-off Hypothesis (e.g., Housen & Kuiken, 2009).

Seven out of the nine participants decreased their speech rate in Week 7, which suggested that the tasks were not equal in difficulty; some storylines might have been more difficult to narrate than others (Sasayama, 2016). As discussed above in regards to the weekly fluctuations in complexity and accuracy, weekly fluctuations in speed fluency might have been influenced by task effects, such as the topic used, the structure of the task, and even picture quality (Elder et al., 2002). Therefore, the weekly equivalent narration task involved a different content and lexis that generated slight differences in task complexity (Sasayama, 2016).

In Week 8 and Week 9, the three FFI + PF group participants increased their speech rate. The increase for speech rate by the FFI + PF group participants in Weeks 8 and 9 might have been due to them reaching a ceiling effect in the performance of percentage of accurate past tense verb usage in Week 7; thus, their syntactic accuracy plateaued in Weeks 8 and 9 and did not further develop. As a consequence, the FFI + PF group participants were able to allocate their attention to developing oral fluency instead. While oral fluency was not traded off with syntactic accuracy at the end of the nine-week period, the FFI + PF group participants' performance of speech rate and percentage of accurate past tense verb usage over Weeks 7, 8 and 9 indicates that there

were week-to-week trade-offs between oral fluency and syntactic accuracy due to limited attentional resources (Bygate, 1998; Ellis, 2005b; Foster & Skehan, 1996).

On the other hand, the form-focused instruction affected speed fluency because the participants did not increase as hypothesized; they should have made gains based on the effects of weekly task repetition. In this regard, form-focused instruction disadvantaged speed fluency because while the intervention group participants improved syntactic accuracy, they did not improve speed fluency more than the comparison group participants.

Form-focused instruction also had no influence on repair fluency as seven participants regardless of group increased in the number of self-repairs. This result was similar to the unanticipated result for Hypothesis 2, where all of the participants declined in repair fluency. Two reasons can be proposed for this decrease in repair fluency. First, significant changes in repair fluency are difficult to identify as previous researchers (e.g., Kormos & Dénes, 2004) have confirmed increases in speed and pause fluency, but not in repair fluency. Repair fluency is therefore considered a less suitable indicator of fluency changes (Lennon, 1990). Second, in the present study, the weekly narration task differed from previous studies in that previous studies were cross-sectional. Therefore, the beneficial effects of repetition in the present study might not have been transferred to the subsequent equivalent task in the following week (Bygate, 2001; Gass et al., 1999). The repair fluency of the participants in the present study might have produced positive results from repetition if the same content had been used for nine sessions or if the study had used a cross-sectional design.

The large week-to-week fluctuations for repair-fluency might have been the result of task complexity, as each week had a different storyline, which required different content and vocabulary. In addition to different content and vocabulary, the equivalent weekly narration task also included different storylines, which might have placed different amounts of pressure on working memory.

Participants who received more form-focused instruction improved more slowly in speed fluency during the research period compared to the other participants. Therefore, while trade-off effects between syntactic complexity, syntactic accuracy, and oral fluency were not present at the end of the research period, they did occur in the participants' week-to-week development.

### Research Hypothesis 4: Differences in Learner Affect

Hypothesis 4 concerned the effects of form-focused instruction on the learners' classroom English speaking anxiety, English speaking self-efficacy, and desire to speak English. Hypothesis 4a stated that the FFI + PF group would increase in classroom English speaking anxiety more than the FFI and comparison groups, and the FFI group would increase more than the comparison group because the FFI + PF group received more form-focused instruction. Hypothesis 4b stated that the FFI + PF group would increase speaking self-efficacy more than the FFI and comparison groups, and the FFI group would increase more than the comparison group because form-focused instruction could increase self-confidence (Van Batenburg et al., 2019). Hypothesis 4c. stated that the FFI + PF group would decrease desire to speak English more than the FFI

and comparison groups, and the FFI group would decrease more than the comparison group because form-focused instruction is likely to decrease learners' willingness to communicate (Van Batenburg et al., 2019).

Three one-way ANCOVAs were conducted to determine the effect of form-focused instruction on learner affect. The results indicated that there was a significant group effect for classroom English speaking anxiety ($p < .05$). Follow-up post hoc tests indicated that the FFI + PF group increased classroom English speaking anxiety significantly more than the FFI group ($p = .03$) and also increased more than the comparison group ($p = .02$). There were no group differences for speaking self-efficacy ($p = .15$), but all three groups increased on the posttest compared to the pretest. For desire to speak English, there was a significant group effect ($p < .05$) and a follow-up post hoc test indicated that the FFI + PF group increased significantly more than the comparison group ($p = .001$). The FFI + PF group also increased more than the FFI group, but the difference was not significant ($p = .26$). The FFI group significantly increased desire to speak English more than the comparison group ($p = .001$).

Analyses of the changes in learner speaking anxiety after the seven-week intervention revealed that form-focused instruction impacted the learners' classroom English speaking anxiety because there was a significant difference between the three groups after the seven-week intervention. The FFI + PF group participants significantly increased classroom English speaking anxiety in comparison to the FFI group participants. It was speculated that the larger amount of form-focused instruction led FFI + PF group participants to see the gaps between their output and the native speaker

model. As a result, they felt that their English was not as good as the FFI group participants who received less FFI; the FFI + PF group participants became more anxious because they were reminded that their output included errors. They then started to focus on producing more accurate forms. While previous studies (e.g., Saito, 2004; Trang, et al., 2013) indicated that teacher instruction can increase anxiety to speak English, there was no difference between the FFI group and the comparison group. Therefore, while excessive amounts of instruction can increase anxiety to speak English in the classroom, smaller amounts of form-focused instruction did not increase learners' anxiety toward speaking English in this study.

Form-focused instruction did not negatively impact speaking self-efficacy because there was no significant difference in the changes in speaking self-efficacy between the three groups after the seven-week intervention. The FFI + PF group participants became more anxious, but they did not decrease speaking self-efficacy. This result refuted previous arguments (Bandura, 1997) that self-efficacy decreases along with increases in anxiety. The participants in all three groups increased speaking self-efficacy, a result that supports Leeming's (2017) results. This result was encouraging, as it addresses the criticism of grammar instruction that it negatively impacts learner affect (Loewen et al., 2009) and supports the argument that some learners prefer explicit form-focused instruction (Graus & Coppen, 2016).

Form-focused instruction possibly influenced the learners' desire to speak English. There was a significant difference in the change of desire to speak English between the three groups after the seven-week intervention. The FFI + PF group

participants significantly increased desire to speak English more than the comparison group participants. This finding was encouraging because a major criticism against form-focused instruction is that grammar instruction demotivates learners (Savignon, 2005) and potentially decreases learners' desire to speak English (Masgoret & Gardner, 2003). It can be speculated that the participants in this study favored direct instruction because it probably made them feel that they were learning something.

### Research Hypothesis 5: Learner Self-Assessment

Hypothesis 5 concerned the learners' self-assessment of their speaking proficiency in terms of self-assessed syntactic complexity, syntactic accuracy, and oral fluency. Hypothesis 5a stated that the FFI group would have higher self-assessment in syntactic accuracy than the comparison group, and the FFI + PF group would have higher self-assessed syntactic accuracy than the FFI and comparison groups because more form-focused instruction should lead to greater syntactic accuracy gains (DeKeyser, 1998; Goh & Burns, 2012; Norris & Ortega, 2000; Spada & Tomita, 2010) and therefore higher self-assessments (Ma & Winke, 2019; Malabonga et al., 2005; Poehner, 2012). Hypothesis 5b stated that the intervention groups will not have higher self-assessment in syntactic complexity nor oral fluency than the comparison group because the form-focused instruction was not designed to produce improvements in those two variables.

Three one-way ANOVAs were conducted to determine the effects of form-focused instruction on learner self-assessed speaking task performance. The ANOVA

results indicated that group effect for syntactic complexity ($p$ = .61), syntactic accuracy ($p$ = .52), and self-assessed oral fluency ($p$ = .29) were not significant. Nevertheless, all three groups increased in all self-assessed syntactic complexity, syntactic accuracy, and fluency after the seven-week intervention.

There were also no significant differences between the three groups in how much they changed in self-assessed syntactic complexity, syntactic accuracy, and oral fluency. The FFI + PF and FFI group participants did not assess themselves as having higher syntactic complexity, syntactic accuracy, or oral fluency than the comparison group participants. One possible reason is because the learners had a different understanding of the purpose and target of the assessment than the instructor (Poehner, 2012). Because the participants were not given self-assessment training, they might have been unable to perceive their own improvement due to a lack of clarity in the assessment benchmarks. High-performing participants might have underestimated their performances while low performing participants might have overestimated theirs (Trofimovich et al., 2016). The effects of form-focused instruction might have been reflected in the learners' self-assessment if they had received self-assessment training prior actual assessment. Previous studies have indicated that learners need to be taught how to assess themselves, as it is a complex process that can only be developed with time, practice, and guidance (Cossar-Burgess & Eberstein, 2013).

While the FFI group participants did not self-assess themselves to have improved more in syntactic accuracy than the comparison group participants, Hypothesis 2 found that that the participants who received form-focused instruction

significantly outperformed those who did not on syntactic accuracy based on CAF measurements. The ANOVAs conducted on the percentage of error-free T-units ($p < .001$, partial $\eta^2 = .15$) and the percentage of accurate past tense verb usage ($p < .001$, partial $\eta^2 = .20$) were significant and post hoc analyses for percentage of error-free T-units indicated that the FFI and FFI + PF groups significantly outperformed the comparison group ($p = .05$). In addition, both the FFI group and FFI + PF group significantly outperformed the comparison group ($p = .05$) in the percentage of accurate past tense verb usage. Many participants who assessed themselves to have not improved in speaking actually did improve. This finding is important because it supports arguments that teachers need to understand how learners assess themselves (Trofimovich et al., 2016) to help them further their language learning.

Finally, while participants who received form-focused instruction did not assess themselves higher than those who did not, all of the participants rated themselves higher in syntactic complexity, syntactic accuracy, and oral fluency after the research period; thus, form-focused instruction had no effect on self-assessment. This finding suggests that teachers need to direct learners to recognize the positive effects of form-focused instruction on speaking proficiency, as the learners need to see gaps in their own learning, specify learning goals, and work to achieve those goals (Little, 2005; Trofimovich et al., 2016). The reflective process of self-assessment not only has the potential of leading learners to recognize the importance of form-focused instruction, but also has other potential positive effects on language learning such as increases in learner motivation and confidence (Gardner, 2000).

## Pedagogical Implications

First, the results indicate that it is desirable to incorporate form-focused instruction into TBLT classrooms because it leads to improvements in syntactic accuracy. Repetition alone does not produce syntactic accuracy improvements; learners will not likely improve syntactic accuracy unless they are explicitly directed to target grammatical features that need improvement. Second, task repetition can be effective because it leads to significant improvements in subordination complexity. Therefore, if the lesson objective is to improve subordination complexity, then integrating repetitive tasks over long periods of time is one effective approach. Third, while form-focused instruction improves syntactic accuracy, it is not detrimental to the development of oral fluency or syntactic complexity. Teachers do not need to be concerned with the potential negative effects of form-focused instruction on speaking proficiency as the trade-off effects between syntactic complexity, syntactic accuracy, and oral fluency did not materialize in this study, notwithstanding such effects in previous studies (e.g., Skehan, 2009a). Nevertheless, teachers need to know that when learners allocate more attention to syntactic accuracy and make improvements to it, they are unlikely to improve oral fluency and/or syntactic complexity. Therefore, in order to improve all three aspects of speaking proficiency, teachers need to provide explicit form-focused instruction as well as different tasks such as pre-task planning and task repetition to improve the other aspects of speaking proficiency. Fourth, while form-focused instruction is effective, excessive amounts of form-focused instruction might not have strong beneficial effects because the participants in both the FFI and FFI + PF groups

improved syntactic accuracy. In addition, learners who receive a larger amount of form-focused instruction might be negatively impacted given that the FFI + PF group participants did not improve speed fluency compared to the FFI group participants who received a smaller amount of form-focused instruction.

The results for Hypothesis 2 further indicated that form-focused instruction can be effectively integrated into TBLT classrooms because it effectively improves learners' syntactic accuracy without negatively impacting syntactic complexity or oral fluency. This finding is encouraging, as previous researchers have found that increases in syntactic accuracy occur at the expensive of decreases in syntactic complexity and/or oral fluency due to trade-off effects. Research on communicative language teaching has generally not been focused on the effects of explicit instruction. Communicative language teaching often overlooks the need to explicitly teach target grammatical forms to learners. As a result, learners might proceduralize their errors, which might lead to the fossilization of these errors (Thai & Boers, 2016). Focus-on-form instruction in this sense is effective as teachers can explicitly teach grammatical knowledge to students by directing the students' attention to target grammatical forms. On the other hand, receiving form-focused instruction from the teacher also encourages learners to pay attention to their own grammatical errors and to consequently produce more accurate target grammatical forms. The integration of form-focused instruction into communication-based tasks supports Swain's (1985) Comprehensible Output Hypothesis, which states that learners develop by first noticing inaccurate grammatical

forms in their own speech, and then they modify their spoken language to achieve greater syntactic accuracy.

Second, despite the importance of form-focused instruction for developing speaking proficiency, teachers need to implement it cautiously, as more instruction is not necessarily more beneficial for learners. As discussed above, form-focused instruction effectively improved the use of past tense in the learners' oral production, but the increase in accurate forms did not require a large amount of instruction time. Therefore, it is important for teachers to adjust the amount of instruction given to students based on the difficulty of the target form after understanding the proficiency and learning background of the learners.

Third, rater judgements of learners' oral performances are less sensitive than CAF measurements (Nitta & Nakatsuhara, 2014). In addition, learners' improvement in the accurate use of simple past tense does not contribute to their analytically rated oral performances because raters might not perceive small improvements in syntactic accuracy. Therefore, teachers need to interpret the results of analytical ratings with caution given that different types of measurement can determine the significance of research results.

Fourth, the results of Hypothesis 3 suggested that while form-focused instruction improves syntactic accuracy, the effectiveness of instruction is likely to decrease gradually as time passes and eventually diminishes entirely when learners reach peak level performance. Therefore, when giving form-focused instruction, it is

useful to target the same linguistic features repeatedly; the learners continued to improve global syntactic accuracy until about Week 7.

Fifth, although increasing the amount of instruction given to the learners did not increase their syntactic accuracy, the difference in the amount of instruction influenced syntactic accuracy development when the interventions were given over one to three weeks. More instruction was more effective between Week 1 and Week 3; however, the effectiveness of more instruction disappeared as time passed. The greater the amount of form-focused instruction provided, the faster the learners developed in the targeted grammatical features; however, once the learners reached a ceiling, the differences between the different amounts of instruction disappeared. When trying to improve specific grammatical features such as tense within three weeks, a greater amount of form-focused instruction is more effective.

Third, teachers need to recognize that syntactic complexity, especially subordination complexity, is difficult to develop. In addition, to develop syntactic complexity, teachers need to provide suitable input, as the form-focused instruction delivered in this study had no positive effects on complexity development. Weekly narration tasks develop the easier global aspect of complexity but not the more complex subordination component. Therefore, teachers need to incorporate other interventions to develop subordination complexity in learners.

Task repetition alone might be sufficient for improving the speed component of speaking fluency. Learners naturally produce self-repairs regardless of whether they receive form-focused instruction or not. Moreover, trade-off effects between repair

fluency and syntactic accuracy do not always occur; therefore, teachers do not need to be overly concerned about the implementation of form-focused instruction in the classroom as trade-off effects do not always take place.

The results of Hypothesis 4 suggested that teachers should include form-focused instruction in speaking classrooms because not only does it not increase learner anxiety to speak English, it also increases their desire to speak the language. In addition, providing form-focused instruction does not negatively affect speaking self-efficacy. However, teachers need to be aware that more grammar instruction is not always better. Similar to the implications suggested by the results of Hypotheses 2 and 3, an excessive amount of form-focused instruction can have negative influences on learner affect. While form-focused instruction positively increased speaking self-efficacy and can potentially improve desire to speak English, it did not cause increases in the classroom English speaking anxiety in cases when form-focused instruction was not over used. Therefore, it can be considered that form-focused instruction is not cause for concern where learner affect is concerned.

The results of Hypothesis 5 suggested that learners' perceptions of the effects of form-focused instruction differ from teacher perceptions; learners are less likely to recognize the importance of form-focused instruction on speaking proficiency development. In addition, it is difficult for learners to notice growth in their own language proficiency. Teachers should not expect learners to be able to effectively conduct self-assessment without giving them adequate training and guidance. Previous researchers (e.g., Little, 2005; Poehner, 2012) have reported that learners who have only

experienced traditional teacher-led assessments cannot accurately assess themselves. Furthermore, without adequate prior training and experience, learner self-assessments can be misleading; thus, they need to be interpreted with caution (Little, 2005). Teachers need to give learners clear and specific benchmarks for conducting self-assessments (Peirce et al., 1993) and learners need to understand the purposes, contents, method of learning, as well as target learning outcomes for conducting their own self-assessment effectively (Poehner, 2012).

Students need training and previous experience to effectively conduct self-assessments (Adams & King, 1995; Jafapur, 1991; Little, 2005) as the amount of prior experience with self-assessment determines the reliability of assessment results (Ross, 1998). While many teachers expect learners to have the ability to appropriately evaluate their own performances, learners who have only experienced traditional teacher assessments cannot self-assess themselves accurately (Little, 2005; Poehner, 2012) and learner self-assessments without teacher mediation can produce unreliable results (Little, 2005). Self-assessment is a complex process that requires learners to develop specific skills and strategies that can only be developed with time, practice, and guidance (Cossar-Burgess & Eberstein, 2013). First, the learners need to understand the purposes, contents, and method of learning as well as target learning outcomes (Poehner, 2012). Second, the learners need unambiguous judgement criteria (Carter & Dunning, 2008) and clear benchmarks (Peirce et al., 1993). Third, the learners need to be competent in the target skill, as incompetent learners might not be able to detect their own incompetence (Carter & Dunning, 2008; Kruger & Dunning, 1999).

One significant finding was that the FFI group and the FFI + PF group participants perceived no improvement in their own task performance despite their significant improvements on some CAF measures. This finding supports arguments that teachers need to understand how learners self-assess their task performance in order to help them further their language learning. While there were no differences between the three groups, all participants assessed themselves slightly higher in syntactic complexity, syntactic accuracy, and fluency after the seven-week intervention.

# CHAPTER 7

# CONCLUSION

In this chapter, I summarize the study and present the main findings for the five research hypotheses, discuss the limitations of the present study, propose suggestions for future research, and make final comments concerning the integration of form-focused instruction in task-based classrooms and its value for the field of second language learning and teaching. In this longitudinal study I examined the effects of form-focused instruction on speaking proficiency development by researching changes in syntactic complexity, syntactic accuracy, and oral fluency of 104 first-year Japanese university students. The study was conducted during English classes during a seven-week intervention, including one pretest, seven-week intervention, and posttest. Participants were divided into three groups: form-focused instruction (FFI), form-focused instruction and peer feedback (FFI + PF), and comparison. The effects of form-focused instruction on classroom English speaking anxiety, English speaking self-efficacy, and desire to speak English were examined as well as speaking proficiency development based on learner self-assessment. Furthermore, the trajectory of weekly changes in syntactic complexity, syntactic accuracy, and oral fluency of nine representative participants from the three groups were also examined.

## Summary of the Findings

First, the results of this study indicated that only participants who received form-focused instruction interventions improved in syntactic accuracy because form-focused instruction led participants to notice the target forms and their attention was directed to produce accurate and precise target forms. Both intervention groups showed no improvements in syntactic complexity possibly because their limited attentional resources were directed at the development of syntactic accuracy. Comparison group participants who received no form-focused instruction improved the global aspect of syntactic complexity slightly; therefore, task repetition might have led to improved syntactic complexity in the case in which the speakers did not have to allocate attentional resources to focusing attention on improving syntactic accuracy. The participants in all three groups demonstrated significant improvements in speed fluency, but repair fluency remained unchanged. The weekly task repetition significantly improved speed fluency because it freed up capacity for retrieving and using new linguistic knowledge, which speeded up language processing and resulted in more automatized speech. The comparison group participants and the FFI group participants significantly improved fluency while the FFI + PF group participants did not. The FFI + PF participants allocated more attention to improving accuracy due to receiving a greater amount of instruction. As a result, they did not significantly improve oral fluency. The results of this study suggest that it cannot be assumed that there are always trade-off effects between accuracy and syntactic complexity even when form-focused instruction is utilized to *focus on accuracy*. This idea differs from researchers (e.g.,

Skehan, 2009b) who have suggested that trade-off effects between accuracy and complexity occur during the process of language learning.

Second, form-focused instruction effectively improved syntactic accuracy. The participants in the FFI and FFI + PF groups who received seven weeks of form-focused instruction significantly outperformed the comparison group participants. Form-focused instruction directs learners' attention to specific learning objectives and encourages the learners to produce more accurate grammatical forms by making cognitive comparisons between their oral output and their explicit knowledge of the target grammatical forms. More instruction might be needed for developing complex grammatical forms such as the past perfect tense, but relatively simple grammatical forms such as past tense can be acquired without spending a great deal of classroom time.

Third, more instruction increased syntactic accuracy more quickly than less instruction. The FFI and FFI + PF group participants improved syntactic accuracy rapidly until they reached a ceiling effect. Syntactic complexity changed very little week-to-week. Form-focused instruction did not have an impact on oral fluency development. Some weekly trade-off effects occurred between syntactic complexity, syntactic accuracy and oral fluency, possibly due to effects of task complexity or the dynamic and non-linear nature of L2 development.

Fourth, limited amounts of form-focused instruction did not increase the participants' classroom English speaking anxiety, but it did increase their desire to speak English. All participants increased English speaking self-efficacy.

Fifth, although there was no significant difference between the three groups based on learner self-assessment, all of the participants assessed themselves slightly higher in syntactic complexity, syntactic accuracy, and fluency after seven weeks of intervention.

## Limitations of the Study

The first limitation is that this study only investigated the effectiveness of instruction of the simple past tense form. The past continuous and past perfect forms were excluded from the analysis because the frequency of their use was less than 1% for past perfect tense and less than 5% for the past continuous form on the posttest. It would have been informative to investigate the effectiveness of form-focused instruction on all three past tense forms.

A second limitation is that this study included different amounts of time on task. The FFI group received 10 minutes of instruction that included a grammar exercise sheet and pair verbal practice, while the FFI + PF group received 10 minutes of instruction plus 10 minutes of peer corrective feedback. The different amount of time on task might have affected the participants in the two intervention groups differently.

The third limitation is that the recordings had different lengths. Although the participants were given two-minutes to conduct their narration and were asked to continue speaking for two minutes, some of the participants stopped their narrations before the two minutes had passed. The different lengths of the recordings could have affected the measurements of fluency, accuracy, and complexity. The participants who

spoke for a longer time likely produced more content than the participants who spoke for a shorter length of time.

Fourth, both rater analytical scoring and self-assessment of the narrative task only included rating criteria regarding syntactic complexity, syntactic accuracy, and oral fluency; items concerning other aspects of speaking such as content organization and pronunciation were not included. The complexity, accuracy, and fluency triad measurement alone has been criticized for its inability to completely grasp the full nature of human speaking proficiency, so including other aspect of speaking proficiency might have led to somewhat different findings (Skehan, 2009b).

Fifth, the participants did not receive training regarding how to conduct self-assessments before engaging in the narrative tasks. As suggested by previous researchers (Cossar-Burgess & Eberstein, 2013), learners need mediation and guidance to perform self-assessment accurately.

The sixth limitation is that the random assignment of the participants to experimental groups was not possible, so the limitations inherent in quasi-experimental studies were unavoidable.

Seventh, the same cartoon was used for the pretest and posttest. This might partially explain the growth in the CAF measures, as some participants commented in the informal interviews that they could remember the storyline in the pretest narrative task. Adding a second narrative task that the participants had not seen previously would have been one way to address this limitation.

Eighth, speech rate in this study was measured after pruning self-repairs and filler pauses. This change can deflate the results of the fluency measures, as would deleting the filler pauses.

## Suggestions for Future Research

The first suggestion is to replicate the current study over a longer period, such as one academic year. In the current study, significant increases from the effect of form-focused instruction during and after seven-week intervention were found for syntactic accuracy, but no significant changes were found in syntactic complexity and oral fluency measurements. Replicating this study over an extended time frame would make it possible to examine whether form-focused instruction impacts the development of syntactic complexity and further explore the dynamic and non-linear nature of L2 speaking development.

Second, it would be useful to examine the effects of instruction on learners at different proficiency levels. It remains uncertain whether the same effects of form-focused instruction apply to learners with higher and lower proficiency levels. Future researchers can consider conducting a study with groups of advanced, intermediate, and beginner learners in order to compare the effects of form-focused instruction on changes in speaking proficiency with such learners.

Third, it would be useful to investigate the effects of different types of form-focused instruction on learners' speaking proficiency. The present study used blackboard teaching, oral and writing grammar exercises, and peer corrective feedback

as form-focused interventions. For instance, peer feedback, and how to best implement it, has garnered significant debate in the SLA literature. Therefore, the investigation of the effects of peer corrective feedback on the development of speaking proficiency in terms of syntactic complexity, syntactic accuracy, and oral fluency could be investigated further.

The fourth suggestions is that future researchers can examine the trajectory of changes in speaking proficiency through qualitative examinations of the language produced in the various tasks. This study used a limited amount of qualitative analyses. Interview data and student learning diary data could have been gathered from the participants. More insight into the development of speaking proficiency can be gained by gathering qualitative data.

Fifth, future researchers can further examine the psychological effects of form-focused instruction on learners. To date, only one other study has been conducted on the effects of form-focused instruction on learner affect (Van Batenburg et al., 2019). The amount of form-focused instruction that is acceptable to learners before their desire to speak declines, if at all, remains unknown. In addition, it is uncertain as to what type of form-focused instruction is perceived positively by learners at different proficiency levels, and who have different English-speaking experiences. Although speaking proficiency is influenced by many factors, including proficiency level and the quality of instruction, it is important for future researchers to examine the psychological effects of different amounts and types of form-focused instruction on learners, as affective variables exert an influence on the development of speaking proficiency.

The sixth suggestion is for researchers to examine the transferability of the effects of language-focused instruction. While Long (2007) suggested that transfer of learning is an objective for every research field of SLA including the field of TBLT, researchers have found little evidence the transfer effects of form-focused instruction. Transfer is more likely to occur with similar tasks, and tasks used in similar situations and settings (Barnett & Ceci, 2002). Therefore, it would be worthwhile for future researchers to add a delayed posttest to examine whether participants can produce target grammatical features in a new context and after a specific period of time.

## Final Conclusions

This study is characterized by its longitudinal design, its large *N*-size of 104 participants, and a detailed description of the week-to-week speaking development of nine participants. In this study, I examined the value of integrating explicit form-focused instruction into meaning centered task-based classrooms by investigating the effects of form-focused instruction on learner speaking proficiency development. Form-focused instruction led to significant improvements in syntactic accuracy, but not syntactic complexity or oral fluency. However, while form-focused instruction can promote the development of syntactic accuracy, more form-focused instruction was not more effective than less instruction. Trade-off effects between syntactic complexity, syntactic accuracy, and oral fluency from the effects of form-focused instruction intervention did not occur at the end of the research period, but some trade-off effects between syntactic accuracy and oral fluency occurred week-to-week during the research

period. Form-focused instruction was effective at developing specific linguistic features more quickly but this effect wore off as time passed. Additionally, it was found that learner self-assessment cannot be used to preciously measure changes in speaking proficiency from the effects of form-focused instruction. Finally, form-focused instruction had minor influences on learners' classroom English speaking anxiety, speaking self-efficacy, and desire to speak English. The present study offers an alternative perspective for examining the development of speaking proficiency that goes beyond the current theoretical boundaries of task-based language teaching and it provides greater epistemological diversity.

## REFERENCES


Adams, M. L. (1980). Five co-occurring factors in speaking proficiency. In J. Firth (Ed.), *Measuring spoken proficiency* (pp. 1-6). Washington, DC: Georgetown University Press.

Adams, C., & King, K. (1995). Towards a framework for student self-assessment. *Innovation in Education and Training International*, *32*(4), 336-343. doi:10.1080/1355800950320405

Ahmadian, M. (2011). The effect of 'massed' task repetitions on complexity, accuracy and fluency: Does it transfer to a new task? *The Language Learning Journal*, *39*(3), 269-280. doi:10.1080/09571736.2010.545239

Ahmadian, M. (2012). The effects of guided careful online planning on complexity, accuracy and fluency in intermediate EFL learners' oral production: The case of English articles. *Language Teaching Research*, *16*(1), 129-149. doi:10.1177/1362168811425433

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

Aida, Y. (1994). Examination of Horwitz, Horwitz, and Cope's construct of foreign language anxiety: The case of students of Japanese. *Modern Language Journal*, *78*(2), 155-168. doi:10.1111/j.1540-4781.1994.tb02026.x

Akao, F. (2011). *Eiken junikyu nijishiken mensetsu kanzen yosou mondai*. [*The Eiken pre-Level 1 study book for the secondary interview test*]. Tokyo, Japan: Obunsha.

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment.* New York, NY: Continuum.

Alderson, J. C., & Banerjee, J. (2001). Language testing and assessment (Part 1). *Language Teaching*, *34*(4), 213-236. doi:10.1017/S0261444800014464

Apple, M. T. (2011). *The big five personality traits and foreign language speaking confidence among Japanese EFL students* (Doctoral dissertation). Available from ProQuest Dissertations & Theses database (3457819). Retrieved from https://search-proquest- com.libproxy.temple.edu/pqdt/docview/874271493/FC5C1BA3039417EPQ/1?accountid=14270

Babaii, E., Taghaddomi, S., & Pashmforoosh, R. (2016). Speaking self-assessment: Mismatches between learner's and teachers' criteria. *Language Testing*, *33*(3), 411-437. doi:10.1177/0265532215590847

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing*, *12*(2), 238-257. doi:10.1177/026553229501200206

Baleghizadeh, S., & Shahri, M. N. N. (2017). The effect of online planning, strategic planning and rehearsal across two proficiency levels. *The Language Learning Journal*, *45*(2), 171-184. doi:10.1080/09571736.2013.808258

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, *84*(2), 191-215. doi:10.1016/0146-6402(78)90002-4

Bandura, A. (1988). Self-efficacy conception of anxiety. *Anxiety Research*, *1*(2), 77-98. doi:10.1080/10615808808248222

Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York, NY: Freeman.

Bardovi-Harlig, K. (2000). *Tense and aspect in second language acquisition: Form, meaning, and use.* Oxford, England: Blackwell.

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin*, *128*(4), 612-37. doi:10.1037/0033-2909.128.4.612

Batstone, R. (1996). Noticing. *ELT Journal*, *50*(3), 273. doi.org/10.1093/elt/50.3.273

Bond, T., & Fox, C. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Erlbaum.

Borg, S. (2003). Teacher cognition in language teaching: A review of research on what language teachers think, know, believe, and do. *Language Teaching*, *36*(2), 81-109. doi:10.1017/S0261444803001903

Brantmeier, C., Vanderplank, R., & Strube, M. (2012). What about me? Individual self-assessment by skill and level of language instruction. *System*, *40*(1), 144-160. doi:10.1016/j.system.2012.01.003

Brindley, G. (2001). Investigating rater consistency in competency-based language assessment. In G. Brindley & C. Burrows (Eds.), *Studies in immigrant English language assessment* (pp. 59-80). Sydney, Australia: National Centre for English Language Teaching and Research, Macquarie University.

Brindley, G., & Slatyer, H. (2002). Exploring task difficulty in ESL listening assessment. *Language Testing, 19*(4), 369-394. doi:10.1191/0265532202lt236oa

Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing*, *12*(1), 1-15. doi:10.1177/026553229501200101

Brown, A. (2000). An investigation of the rating process in the IELTS oral interview. In R. Tulloh (Ed.), *International English Language Testing System (IELTS) Research Reports 2000, Volume 3* (pp. 49-84). Melbourne: IELTS Australia. Retrieved from https://www.ielts.org/-/media/research-reports/ielts_rr_volume03_report3.ashx

Brown, J. D., & Bailey, K. M. (1984). A categorical instrument for scoring second language writing skills. *Language Learning*, *34*(4), 21-38. doi:10.1111/j.1467-1770.1984.tb00350.x

Brown, J. D., & Hudson, T. (1998). The alternatives in language assessment. *TESOL Quarterly, 32*(4), 653-675. doi:10.2307/3587999

Brumfit, C. J. (1984). *Communicative methodology in language teaching*. Cambridge, England: Cambridge University Press.

Butler, Y. G. (2011). The Implementation of communicative and task-based language teaching in the Asia-Pacific region. *Annual Review of Applied Linguistics, 31,* 36-57. doi:10.1017/S0267190511000122

Bygate, M. (1996). Effects of task repetition: Appraising the developing language of learners. In J. Willis & E. Willis (Eds.), *Challenge and change in language teaching* (pp. 136-146). Oxford, England: Macmillan.

Bygate, M. (1998). Theoretical perspectives on speaking. *Annual Review of Applied Linguistics, 18*, 20-42.

Bygate, M. (1999). Quality of language and purpose of task: Patterns of learners' language on two oral communication tasks. *Language Teaching Research, 3*(3), 185-214. doi:10.1177/136216889900300302

Bygate, M. (2001). Effects of task repetition on the structure and control of language. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks, second language learning, teaching and testing* (pp. 23-48). London, England: Routledge.

Bygate, M. (2009). Teaching and testing speaking. In C. Doughty & M. H. Long (Eds.), *Handbook of second and foreign language teaching* (pp. 412-440). New York, NY: Blackwell.

Bygate, M., & Samuda, V. (2005). Integrative planning through the use of task repetition. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 37-74). Amsterdam, The Netherlands: Benjamins.

Bygate, M., Skehan, P., & Swain, M. (2001). Introduction. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks, second language learning, teaching and testing* (pp. 1-21). London, England: Routledge.

Cai, H. (2015). Weight-based classification of raters and rater cognition in an EFL speaking test. *Language Assessment Quarterly*, *12*(3), 262-282. doi:10.1080/15434303.2015.1053134

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, *1*(1), 1-47. doi:10.1093/applin/I.1.1

Carter, T. J., & Dunning, D. (2008). Faulty self-assessment: Why evaluating one's own competence is an intrinsically difficult task. *Social and Personality Psychology Compass*, *2*(1), 346-360. doi:0.1111/j.1751-9004.2007.00031.x

Celce-Murcia, M., & Olshtain, E. (2000). *Discourse and context in language learning*. Cambridge, England: Cambridge University Press.

Cheng, Y., Horwitz, E. K., & Schallert, D. L. (1999). Language anxiety: Differentiating writing and speaking components. *Language Learning*, *49*(3), 417-446. doi:10.1111/0023-8333.00095

Clément, R., Dörnyei, Z., & Noels, K. A. (1994) Motivation, self-confidence, and group cohesion in the foreign language classroom. *Language Learning*, *44*(3), 417-448. doi:10.1111/j.1467-1770.1994.tb01113.x

Cossar-Burgess, D., & Eberstein, A. (2013). Encouraging students to become independent learners through self-assessment and reflection. *Cambridge English Research Notes*, *53*, 18-26.

Costa, P. (2015). Reenvisioning language anxiety in the globalized classroom through a social imaginary lens. *Language Learning*, *65*(3), 504-532. doi:10.1111/lang.12121

Davis, L. (2016). The influence of training and experience on rater performance in scoring spoken language. *Language Testing*, *33*(1), 117-135. doi:10.1177/0265532215582282

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

Dell, G. S. (1986). A spreading activation theory of retrieval in sentence production. *Psychological Review*, *93*(3), 283-321. doi:10.1037/0033-295X.93.3.283

de Jong, N. H. (2016). Fluency in second language assessment. In D. Tsagari & J. Banerjee (Eds.), *Handbook of second language assessment* (pp. 203-218). Boston, MA: Mouton de Gruyter.

de Jong, N. H., Groenhout, R., Schoonen, R., & Hulstijn, J. H. (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. H., & Vercellotti, M. L. (2016). Similar prompts may not be similar in the performance they elicit: Examining fluency, complexity, accuracy, and lexis in narrative form five picture prompts. *Language Teaching Research*, *20*(3), 387-404. doi:10.1177/1362168815606161

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). Cambridge, England: 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.

Dell, G. S. (1986). A spreading activation theory of retrieval in sentence production. *Psychological Review*, *93*(3), 283-321. doi:10.1037/0033-295X.93.3.283

Dewaele, J. M., & MacIntyre, P. D. (2014). The two faces of Janus? Anxiety and enjoyment in the foreign language classroom. *Studies in Second Language Learning and Teaching*, *4*(2), 237–274. doi:10.14746/ssllt.2014.4.2.5

Dolean, D. D. (2016). The effects of teaching songs during foreign language classes on students' foreign language anxiety. *Language Teaching Research*, *20*(5), 638-653. doi:10.1177/1362168815606151

Dörnyei, Z. (2003). *Questionnaires in second language research construction, administration, and processing*. Mahwah, NJ: Erlbaum.

Dörnyei, Z. (2005). *The psychology of the language learner: Individual differences in second language acquisition*. Mahwah, NJ: Erlbaum.

Doughty, C. (2001). Cognitive underpinnings of focus on form. In P. Robinson (Ed.), *Cognitive and second language instruction* (pp. 206-257). Cambridge, England: Cambridge University Press.

Doughty, C., & Williams, J. (1998). Issues and terminology. In C. Doughty & J. Williams (Eds.), *Focus on form in classroom second language acquisition* (pp. 1-11). Cambridge, England: Cambridge University Press.

Doughty, C. & Varela, E. (1998). Communicative focus on form. In C. Doughty & J. Williams (Eds.), *Focus on form in classroom second language acquisition* (pp. 114-138). Cambridge, England: Cambridge University Press.

Elder, C., & Iwashita, N., & McNamara, T. (2002). Estimating the difficult of oral proficiency tasks: What does the test-taker have to offer? *Language Testing*, *19*(4), 347-368. doi:10.1191/0265532202lt235oa

Elder, C., & Iwashita, N. (2005). Planning for test performance: Does it make a difference? In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 219-238). Amsterdam, The Netherlands: Benjamins.

Ellis R. (1995). Interpretation tasks for grammar teaching. *TESOL Quarterly*, *29*(1), 87-105. doi:10.2307/3587806

Ellis R. (2001). Introduction: Investigating form-focused instruction. *Language Learning*, *51*(s1), 1-46. doi:10.1111/j.1467-1770.2001.tb00013.x

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

Ellis, R. (2005a). *Analysing learner language*. New York, NY: Oxford University Press.

Ellis, R. (2005b). Instructed language learning and task-based learning. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 713-727). New York, NY: Routledge.

Ellis, R. (2006). Current issues in the teaching of grammar: An SLA perspective. *TESOL Quarterly*, *40*(1), 83-107. doi:10.2307/40264512

Ellis, R. (2009a). 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. (2009b). 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. (2013). Task-based language teaching: Responding to the critics. *University of Sydney Papers in TESOL, 8*, 1-17.

Ellis, R. (2018). Reflections on task-based language teaching. Bristol, England: Multilingual Matters.

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

Ellis, R., Basturkmen, H., & Loewen, S. (2001). Learner uptake in communicative ESL lessons. *Language Learning*, *51*(2), 281-318. doi:10.1111/1467-9922.00156

Ellis, R., Basturkmen, H., & Loewen, S. (2001). Preemptive focus on form in the ESL classroom. *TESOL Quarterly*, *35*(3), 407-432. doi:10.2307/3588029

Ellis, R., Loewen, S., & Erlam, R. (2006). Implicit and explicit corrective feedback and the acquisition of L2 grammar. *Studies in Second Language Acquisition*, *28*(2), 339-368. doi:10.1017/S0272263106060141

Ellis, R., & Yuan, F. (2004). The effects of planning on fluency, complexity, and accuracy in second language narrative writing. *Studies in Second Language Acquisition*, *26*(1), 59-84. doi:10.1017/S0272263104026130

Ewald, J. D. (2007). Foreign language learning anxiety in upper-level classes: Involving students as researchers. *Foreign Language Annals*, *40*(1), 122-142. doi.org/10.1111/j.1944-9720.2007.tb02857.x

Fan, J., & Bond, T. (2019). Applying Rasch measurement in language assessment. In. V. Aryadoust & M. Raquel (Eds.), *Quantitative data analysis for language assessment* (pp. 83-102). London, England: Routledge.

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

Fisher, W. P. Jr. (1992). Reliability, separation, strata statistics. *Rasch Measurement Transactions*, *6*(3), 238. Retrieved from http://rasch.org/rmt/rmt63i.htm

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

Foster, P., & Skehan, P. (1996). The influence of planning and task type on second language performance. *Studies in Second Language Acquisition*, *18*(3), 299-323. doi:10.1017/S0272263100015047

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

Fotos, S. (2002). Structure-based interactive tasks for the EFL grammar learner. In E. Hinkel & S. Fotos (Eds.), *New perspectives on grammar teaching in second language classrooms* (pp. 135-154). Mahwah, NJ: Erlbaum.

Freed, B. F. (2000). Is fluency, like beauty, the eyes, of the beholder? In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 243-265). Ann Arbor, MI: The University of Michigan Press.

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

Gan, Z. (2013). Understanding English speaking difficulties: An investing of two Chinese populations. *Journal of Multilingual and Multicultural Development*. *34*(3), 231-248. doi:10.1080/01434632.2013.768622

Gardner, R. C. (1985). *Social psychology and second language learning: The role of attitudes and motivation*. London, England: Edward Arnold.

Gardner, D. (2000). Self-assessment for autonomous language learners, *Links and Letters*, *7*, 49-60.

Gardner, R. C., Tremblay, P. F., & Masgoret, A. (1997). Towards a full model of second language learning: An empirical investigation, *The Modern Language Journal*, *31*(3), 344-362. doi:10.1111/j.1540-4781.1997.tb05495.x

Gass, S., Mackey, A., Alvarez-Torres, M., & Fernández-García, M. (1999). The effects of task repetition on linguistic output. *Language Learning*, *49*(4), 549-580. doi:10.1111/0023-8333.00102

George, D., & Mallery, P. (2016). *IBM SPSS statistics 23 step by step a simple guide and reference*. New York, NY: Routledge.

Gkonou, C. (2014). The sociolinguistic parameters of L2 speaking anxiety. *Second Language Learning and Teaching*, *17*, 15-32. doi:10.1007/978-3-319-00188-3_2

Goh, C. C. M., & Burns, A. (2012). *Teaching speaking: A holistic approach*. New York, NY: Cambridge University Press.

Graus, J., & Coppen, P. (2016). Student teacher beliefs on grammar instruction. *Language Teaching Research*, *20*(5), 571-599. doi:10.1177/1362168815603237

Green, S. B., & Salkind, N. J. (2011). *Using SPSS for Windows and Macintosh: Analyzing and understanding data* (6th ed.). Boston, MA: Prentice Hall.

Gregersen, T. S., & Horwitz, E. (2002). Language learning and perfectionism: Anxious and non-anxious language learners' reactions to their own oral performance. *The Modern Language Journal*, *86*(4), 562-570. doi:10.1111/1540-4781.00161

Guénette, D. (2007). Is feedback pedagogically correct? Research design issues in studies of feedback on writing. *Journal of Second Language Writing*, *16*(1), 40-53. doi:10.1016/j.jslw.2007.01.001

Halliday, M. A. K., & Matthiessen, C. (1999). *Construing experience through meaning: A language-based approach to cognition*. London, England: Cassell.

Hilton, H. (2008). The link between vocabulary knowledge and spoken L2 fluency. *The Language Learning Journal*, *36*(2), 153-166. doi:10.1080/09571730802389983

Han, Z. H. (2004). Fossilization in adult second language acquisition. Clevedon, England: Multilingual Matters.

Horwitz, E. K. (2010). Foreign and second language anxiety. *Language Teaching*, *43*(2), 154-167. doi:10.1017/S026144480999036X

Horwitz, E. K., Horwitz, M. B., & Cope, J. (1986). Foreign language classroom anxiety. *The Modern Language Journal*, *70*(2), 125-132. doi:10.1111/j.1540-4781.1986.tb05256.x

Horwitz, E., & Young, D. (1991). *Language anxiety: From theory and research to classroom implications*. Englewood Cliffs, NJ: Prentice Hall.

Housen, A., & Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics*, *30*(4), 461-473. doi:10.1093/applin/amp048.

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

Hughes, R. (2011). *Teaching and researching speaking*. New York, NY: Routledge.

Hunt, K. W. (1970). Syntactic maturity in school children and adults. *Monographs of the Society for Research in Child Development*, *35*(1) (serial no. 134), 1-67.

Incecay, V., & Dollar, Y. K. (2011). A foreign language learner's beliefs about grammar instruction and error correction. *Procedia Social and Behavioral Sciences*, *15*, 3394-3398. doi:10.1016/j.sbspro.2011.04.307

Iwashita, N. (2010). Features of oral proficiency in task performance by EFL and JFL learners. In M. T. Prior, Y. Watanabe, & S. Lee (Eds.), *Selected proceedings of the 2008 second language research forum: Exploring SLA perspectives, positions, and practices* (pp. 32-47). Honolulu, HI: Second Language Research Forum.

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

Iwashita, N., Elder, C., & McNamara, T. (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

Izumi, S. (2002). Output, input enhancement, and the noticing hypothesis: An experimental study on ESL relativization. *Studies in Second Language Acquisition*, *24*(4), 541-577. doi:10.1017/S0272263102004023

Izumi, S. (2003). Comprehension and production processes in second language learning: In search of the psycholinguistic rationale of the output hypothesis, *Applied Linguistics*, *24*(2), 168-196. doi:10.1093/applin/24.2.168

Jafapur, A. (1991). Can naive EFL learners estimate their own proficiency? *Evaluation and Research in Education*, *5*(3), 145-57. doi:10.1080/09500799109533306

Kanda, M. (2015). Development of English oral proficiency among Japanese high school students (Doctoral dissertation). Available from ProQuest Dissertations & Theses Database (1754638039). Retrieved from https://search-proquestcom.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. 143-164). Amsterdam, The Netherlands: Benjamins.

Kitano, K. (2001). Anxiety in the college Japanese language class. *The Modern Language Journal*, *85*(4), 549-566. doi:10.1111/0026-7902.00125

Koizumi, R., & In'nami, Y. (2014). Modeling complexity, accuracy, and fluency of Japanese learners of English: A structural equation modeling approach. *JALT Journal*, *36*, 25-46.

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

Kormos, J. (2006). *Speech production and second language acquisition*. Mahwah, NJ: Erlbaum.

Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System*, *32*(2), 145-164. doi:10.1016/j.system.2004.01.001

Krashen, S. (1981). *Second language acquisition and second language learning*. Oxford, England: Pergamon.

Kramsch, C. (2014). Teaching foreign language in an era of globalization: Introduction. *The Modern Language Journal*, *98*(1), 296-311. doi:10.1111/j.1540-4781.2014.12057.x

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, *77*(6), 1121-1134. doi:10.1037//0022-3514.77.6.1121

Kuiken, F., & Vedder, I. (2014). Raters' decisions, rating procedures and rating scales. *Language Testing*, *31*(3), 279-284. doi:10.1177/0265532214526179

Lam, W. Y. K. (2006). Gauging the effects of ESL oral communication strategy teaching: A multimethod approach. *Electronic Journal of Foreign Language Teaching*, *3*(2), 142-157. Retrieved from http://e-flt.nus.edu.sg/v3n22006/lam.pdf

Lambert, C., & Kormos, J. (2014). Complexity, accuracy, and fluency in task-based L2 research: Toward more developmentally based measures of second language acquisition. *Applied Linguistics*, *35*(5), 607-614. doi:10.1093/applin/amu047

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. (1986). *Techniques and principles in language teaching*. New York, NY: Oxford University Press.

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/apphn/aml029

Larsen-Freeman, D. (2009). Adjusting expectations: The study of complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics, 30*(4), 579-589. doi:10.1093/applin/amp043

Lee, S. C. N., & Tajino, A. (2008). Understanding students' perceptions of difficulty with academic writing for teacher development: A case study of the University of Tokyo writing program. *Kyoto University Researches in Higher Education*, 14, 1-11.

Leeming, P. (2017). A longitudinal investigation into English speaking self-efficacy in a Japanese language classroom. *Asian-Pacific Journal of Second and Foreign Language Education*, *2*(12), 1-18. doi:10.1186/s40862-017-0035-x

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

Lennon, P. (2000). The lexical element in spoken second language fluency. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 25-42). Ann Arbor, MI: The University of Michigan Press.

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

Li, D. (1998). It's always more difficult than you plan and imagine: Teachers' perceived difficulties in introducing the communicative approach in South Korea. *TESOL Quarterly, 32*(4), 677-703. doi:10.2307/3588000

Li, S. F. (2010). The effectiveness of corrective feedback in SLA: A meta-analysis. *Language Learning, 62*(2), 309-365. doi:10.1111/j.1467-9922.2010.00561.x

Linacre, J. M. (2002a). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions, 16*(2), 878. Retrieved from http://www.rasch.org/rmt/rmt162f.htm

Linacre, J. M. (2002b). Optimizing rating scale category effectiveness. *Journal of Applied Measurement, 3*(1), 85-106. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.424.2811&rep=rep1&type=pdf

Linacre, J. M. (2007). *A user's guide to WINSTEPS: Rasch-model computer program*. Chicago, IL: MESA.

Linacre, J. M. (2011). *A user's guide to FACETS: Rasch-model computer program*. Chicago, IL: MESA.

Linacre, J. M. (2013). *Facets computer program for many-facet Rasch measurement* [Computer software]. Beaverton, Oregon: Winsteps.com

Linacre, J. M., & Wright, B. D. (2007). *WINSTEPS: Multiple-choice, rating scale, and partial credit Rasch analysis* [Computer software]. Chicago, IL: MESA.

Ling, G., Mollaun, P., & Xi., X. (2014). A study on the impact of fatigue on human raters when scoring speaking responses. *Language Testing*, *31*(4), 479-499. doi:10.1177/0265532214530699

Little, D. (2005). The Common European Framework and the European language portfolio: Involving learners and their judgments in the assessment process. *Language Testing*, *22*(3), 321-336. doi:10.1191/0265532205lt311oa

Liu, M. (2006). Anxiety in Chinese EFL students at different proficiency levels. *System, 34*(3), 301-316. doi:10.1016/j.system.2006.04.004

Lochland, P. W. (2013). Moving beyond communicative language teaching: A situated pedagogy for Japanese EFL classrooms. *TESOL Journal*, *4*(2), 261-273. doi:10.1002/tesj.57

Loewen, S., Li, S., Fei, F., Thompson, A., Nakatsukasa, K., Ahn, S., & Chen, X. (2009). Second language learner's beliefs about grammar instruction and error correction. *The Modern Language Journal, 93*(1), 91-104. doi:10.1111/j.1540-4781.2009.00830

Long, M. H. (1985). Input and second language acquisition theory. In S. Gass & C. Madden (Eds.), *Input in second language acquisition* (pp. 377-393). Rowley, MA: Newbury House.

Long, M. H. (1991). Focus on form: A design feature in language teaching methodology. In K. de Bot, R. Ginsberg, & C. Kramsch (Eds.), *Foreign language research in cross-cultural perspective* (pp. 39-52). Amsterdam, The Netherlands: Benjamins.

Long, M. H. (2007). SLA: Breaking the siege. In M. H. Long (Ed.), *Problems in SLA* (pp. 139-168). Mahwah, NJ: Erlbaum.

Long, M. H. (2015). *Second language acquisition and task-based language teaching*. West Sussex, England: John Wiley & Sons.

Long, M. H., & Crookes, G. (1992). Three approaches to task-based syllabus design. *TESOL Quarterly*, *26*(1), 27-56. doi:10.2307/3587368

Lynch, T., & Maclean, J. (2000). Exploring the benefits of task repetition and recycling for classroom language learning. *Language Teaching Research*, *4*(3), 221-50.

Lyster, R. (2004). Research on form-focused instruction in immersion classrooms: Implications for theory and practice. *Journal of French Language Studies*, *14*(3), 321-341. doi:10.1017/S0959269504001826

Lyster, R., & Saito, K. (2010). Oral feedback in classroom SLA: A meta-analysis. *Studies in Second Language Acquisition*, *32*(2), 265-302. doi:10.1017/S0272263109990520

Ma, W., & Winke, P. (2019). Self-assessment: How reliable is it in assessing oral proficiency over time? *Foreign Language Annals*, 52(1), 66-86. doi:10.1111/flan.12379

MacIntyre, P. D. (2002). Motivation, anxiety and emotion in second language acquisition. In P. Robinson (Ed.), *Individual differences and instructed language learning, Volume 2* (pp. 45-68). Philadelphia, PA: Benjamins.

MacIntyre, P. D., Babin, P. A., & Clément, R. (1999). Willingness to communicate: Antecedents & consequences. *Communication Quarterly*, *47*(2), 215-229. doi:10.1080/01463379909370135

MacIntyre, P. D., Baker, S. C., Clément, R., & Conrod, S. (2001). Willingness to communicate, social support, and language-learning orientations of immersion students. *Studies in Second Language Acquisition*, *23*(3), 369-388. doi:10.1017/S0272263101003035

MacIntyre, P. D., Baker, S. C., Clément, R., & Donovan, L. A. (2003). Talking in order to learn: Willingness to communicate and intensive language programs. *Canadian Modern Language Review*, *59*(4), 589-608. doi:10.3138/cmlr.59.4.589

MacIntyre, P. D., Dörnyei, Z., Clément, R., & Noels, K. A. (1998). Conceptualizing willingness to communicate in a L2: A situational model of L2 confidence and affiliation. *The Modern Language Journal*, *82*(4), 545-562. doi:10.1111/j.1540-4781.1998.tb05543.x

MacIntyre, P. D., & Gardner, R. C. (1991). Methods and results in the study of anxiety and language learning: A review of the literature. *Language Learning*, *41*(1), 85-117. doi:10.1111/j.1467-1770.1991.tb00677.x

MacIntyre, P. D., Noels, K. A., & Clément, R. (1997). Biases in self-ratings of second language proficiency: The role of language anxiety. *Language Learning*, *47*(2), 265-287. doi.org/10.1111/0023-8333.81997008

Mackey, A. (2006). Feedback, noticing and instructed second language learning. *Applied Linguistics*, *27*(3), 405-530. doi:10.1093/applin/ami051

Mak, B. (2011). An exploration of speaking-in-class anxiety with Chinese ESL learners. *System*, *39*(2), 202-214. doi:10.1016/j.system.2011.04.002

Malabonga, V., Kenyon, D. M., & Carpenter, H. (2005). Self-assessment, preparation and response time on a computerized oral proficiency test. *Language Testing*, *22*(1), 59-92. doi:10.1191/0265532205lt297oa

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

Masgoret, A. M., & Gardner, R. C. (2003). Attitudes, motivation, and second language learning: A meta-analysis of studies conducted by Gardner and associates. *Language Learning*, *53*(1), 123-163. doi:10.1111/1467-9922.00212

Matsuno, S. (2009). Self-, peer- and teacher-assessments in Japanese university EFL writing classrooms. *Language Testing*, *26*(1), 75-100. doi:10.1177/0265532208097337

Mehnert, U. (1998). The effects of different lengths of time for planning on second language performance. *Studies in Second Language Acquisition*, *20*(1), 52-83. doi:10.1017/S0272263198001041

Michel, M. (2017). Complexity, accuracy, and fluency (CAF). In S. Loewen & M. Sato (Eds.), *The Routledge handbook of instructed second language acquisition* (pp. 66-94). London, England: Routledge.

Michel, M. C., Kuiken, F., & Vedder, I. (2007). The influence of complexity in monologic versus dialogic tasks in Dutch L2. *International Review of Applied Linguistics in Language Teaching*, *45*(3), 241-59. doi:10.1515/iral.2007.011

Mills, N. (2014). Self-efficacy in second language acquisition. In S. Mercer & M. Williams (Eds.), *Multiple perspectives on the self in SLA* (pp. 6-19). Bristol, England: Multilingual Matters.

Mills, N., Pajares, F., & Herron, C. (2006). A reevaluation of the role of anxiety: Self-efficacy, anxiety, and their relation to reading and listening proficiency. *Foreign Language Annals*, *39*(2), 276-294. doi:10.1111/j.1944-9720.2006.tb02266.x

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

Mora, J. (2006). Age effects on oral fluency development. In C. Munoz (Ed.), A*ge and the rate of foreign language learning* (pp. 65-88). New York, NY: Multilingual Matters.

Wordcalc. (2019). Syllable counter. Retrieved from https://www.wordcalc.com/contact.php

Mori, Y. (1999). Epistemological beliefs and language learning beliefs: What do language learners believe about their learning? *Language Learning*, *49*(3), 377-415. doi:10.1111/0023-8333.00094

Müge Satar, H., & Özdener, N. (2008). The effects of synchronous CMC on speaking proficiency and anxiety: Text versus voice chat. *The Modern Language Journal*, *92*(4), 595-613. doi:10.1111/j.1540-4781.2008.00789.x

Muranoi, H. (2000). Focus on form through interaction enhancement: Integrating formal instruction into a communicative task in EFL classrooms. *Language Learning*, *50*(4), 617-673. doi:10.1111/0023-8333.00142

Murphy, R. (2012). *English grammar in use: A self-study reference and practice book for intermediate learners of English* (4th ed.). Cambridge, England: Cambridge University Press.

Nassaji, H. (2000). Towards integrating form-focused instruction and communicative interaction in the second language classroom: Some pedagogical possibilities. *The Modern Language Journal*, *84*(2), 241-250. doi:10.1111/0026-7902.00065

Nassaji, H., & Fotos, S. (2011). *Teaching grammar in second language classrooms: Integrating form-focused instruction in communicative context*. 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.

Nicholas, H., Lightbown, P., & Spada, N. (2008). Recasts as feedback to language learners. *Language Learning*, *51*(4), 719-758. doi:10.1111/0023-8333.00172

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

Norris, J. M., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning*, *50*(3), 417-528. doi:10.1111/0023-8333.00136

Norris, J. M., & Ortega, L. (2003). Defining and measuring SLA. In C. J. Doughty & M. H. Long (Eds.), *Handbook of second language acquisition* (pp. 717-761). Malden, MA: Blackwell.

Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics*, *30*(4), 555-578. doi.org/10.1093/applin/amp044

Nunan, D. (1988). *The learner-centered curriculum: A study in second language teaching*. Cambridge, England: Cambridge University Press.

Nunan, D. (1989). *Designing tasks for a communicative classroom*. Cambridge, England: Cambridge University Press.

Nunan, D. (2003). The impact of English as a global language on educational policies and practices in the Asia. *TESOL Quarterly*, *37*(4), 589-613. doi:10.2307/3588214

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., Iwashita, N., Rabie, S., & Norris, J. (1999, August). *A multilanguage comparison of L2 syntactic complexity measures*. Presentation presented at the 12th World Congress of International Association of Applied Linguistics Conference, Tokyo, Japan.

Patri, M. (2002). The influence of peer feedback on self- and peer-assessment of oral skills. *Language Testing*, *19*(2), 109-131. doi:10.1191/0265532202lt224oa

Paulus, T. M. (1999). The effect of peer and teacher feedback on student writing. *Journal of Second Language Writing*, *8*(3), 265-289. doi:10.1016/S1060-3743(99)80117-9

Peirce, B. N., Swain, M., & Hart, D. (1993). Self-assessment, French immersion, and locus of control. *Applied Linguistics*, *14*(1), 25-42. doi:10.1093/applin/14.1.98

Piniel, K., & Csizér, K. (2013). L2 motivation, anxiety and self-efficacy: The interrelationship of individual variables in the secondary school context. *Studies in the Second Language Learning and Teaching*, *3*(4), 523-550. doi:10.14746/ssllt.2013.3.4.5

Poehner, M. E. (2012). The zone of proximal development and the genesis of self-assessment. *The Modern Language Journal*, *96*(4), 610-622. doi:10.1111/j.1540-4781.2012.01393.x

Polio, A. (1997). Measures of linguistic accuracy in second language writing research. *Language Learning*, *47*(1), 101-143. doi:10.1111/0023-8333.31997003

Prabhu, N. S. (1987). *Second language pedagogy*. Oxford, England: Oxford University Press.

Rahimpour, M., Salimi, A., & Farrokhi, F. (2012). The effect of intensive and extensive focus on form on EFL learners' written accuracy. *Theory and Practice in Language Studies, 2*(11), 2277-2283. doi:10.4304/tpls.2.11.2277-2283

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: University of Chicago.

Revesz, A., Sachs, R., & Hama, M. (2014). The effects of task complexity and input frequency on the acquisition of the past counterfactual construction through recasts. *Language Learning*, *64*(3), 615-650. doi:10.1111/lang.12061

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

Riggenbach, H. (1991). Towards an understanding of fluency: A microanalysis of nonnative speaker conversation. *Discourse Processes, 14*(4), 423-441. doi:10.1080/01638539109544795

Robinson, P. (2001). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics, 22*(1), 27-57. doi:10.1093/applin/22.1.27

Robinson, P. (2011). Second language task complexity, the Cognition Hypothesis, language learning, and performance. In Robinson, P. (Ed.), *Second language task complexity: Researching the Cognition Hypothesis of language learning and performance* (pp. 3-38). Amsterdam, The Netherlands: Benjamins.

Ross, S. (1998). Self-assessment in second language testing: A meta-analysis and analysis of experiential factors. *Language Testing, 15*(1), 1-20. doi:10.1177/026553229801500101

Saito, H. (2004). Seeing English language teaching and learning through the eyes of Japanese EFL and ESL students. *Foreign Language Annals, 37*(1), 111-124. doi:10.1111/j.1944-9720.2004.tb02178.x

Saito, H., & Fujita, T. (2004). Characteristics and user acceptance of peer rating in EFL writing classrooms, *Language Teaching Research, 8*(1), 31-54. doi:10.1191/1362168804lr133oa

Sakui, K. (2004). Wearing two pairs of shoes: Language teaching in Japan. *ELT Journal, 58*, 155-163. doi:10.1093/elt/58.2.155

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.

Sasayama, S. (2016). Is a 'complex' task really complex? Validating the assumption of cognitive task complexity. *The Modern Language Journal, 100*(1), 231-254. doi:10.1111/modl.12313

Sato, M. (2014). Explore the construct of interactive fluency: Second language acquisition and language testing approaches. *System, 45,* 79-91. doi:10.1016/j.system.2014.05.004

Sato, T. (2011). The contribution of test-takers' speech content to scores on an English oral proficiency test. *Language Testing, 29*(2), 23-241. doi:10.1177/0265532211421162

Savignon, S. J. (2005). Communicative language teaching: Strategies and goals. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 635-651). New York, NY: Routledge.

Sawaki, Y. (2007). Construct validation of analytic rating scales in a speaking assessment: Reporting a score profile and a composite. *Language Testing, 24*(3), 355-390. doi:10.1177/0265532207077205

Schmidt, R. (1990). The role of consciousness in second language learning. *Applied Linguistics, 11*(2), 129-158. doi.org/10.1093/applin/11.2.129

Schmidt, R. (2001). Attention. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 3-32). Cambridge, England: Cambridge University Press.

Schmidt, R., & Frota, S. N. (1986). Developing basic conversational ability in a second language: A case study of an adult learner of Portuguese. In R. R. Day (Ed.), *Talking to learn: Conversation in second language acquisition* (pp. 237-326). Rowley, MA: Newbury House.

Schunk, D. H. (1991). Self-efficacy and academic motivation. *Educational Psychologist, 26*(3-4), 207-231. doi:10.1207/s15326985ep2603&4_2

Scott, C. M. (1988). Spoken and written syntax. In M. Nippold (Ed.), *Later language development: Ages nine through nineteen* (pp. 49-95). Austin, TX: Pro-Ed.

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

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

Sheppard, C. (2006). *The effects of instruction directed at the gaps second language learners noticed in their oral production*. (Unpublished doctoral dissertation). University of Auckland, Auckland, New Zealand.

Skehan, P. (1989). *Individual differences in second language*. London, England: Edward Arnold.

Skehan, P. (1996). A framework for the implementation of task-based instruction. *Applied Linguistics*, *17*(1), 38-62. doi.org/10.1093/applin/17.1.38

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

Skehan, P. (2003). Task-based instruction. *Language Teaching*, *36*(1), 1-14. doi:10.1017/S026144480200188X

Skehan, P. (2009a). Modeling second language performance: Integrating complexity, accuracy, fluency and lexis. *Applied Linguistics*, *30*(4), 510-532. doi:10.1093/applin/amp047

Skehan, P. (2009b). Lexical performance by native and non-native speakers on language-learning tasks. In B. Richards, H. Daller, D. D. Malvern, & P. Meara (Eds.), *Vocabulary studies in first and second language acquisition: The interface between theory and application* (pp. 107-24). London, England: Palgrave Macmillan.

Skehan, P. (2014). Limited attentional capacity, second language performance, and task-based pedagogy. In P. Skehan (Ed.), *Processing perspectives on task performance* (pp. 211-260). Amsterdam, The Netherlands: Benjamins.

Skehan, P. (2018). *Second language task-based performance: Theory, research, assessment*. New York, NY: Routledge, Taylor & Francis Group.

Skehan, P., Bui, G., & Wang, Z. (2012). The task is not enough: Processing approaches to task-based performance. *Language Teaching Research*, *16*(2), 170-187. doi:10.1177/1362168811428414

Skehan, P., & Foster, P. (1999). The influence of task structure and processing conditions on narrative retellings. *Language Learning*, *49*(1), 93-120. doi.org/10.1111/1467-9922.00071

Skehan, P., & Foster, P. (2007). Complexity, accuracy, fluency and lexis in task-based performance: A meta-analysis of the Ealing research. In P. Van Daele, S. Housen, F. Kuiken, M. Pierrard, & I. Vedder (Eds.), *Complexity, accuracy and fluency in second language use, learning and teaching* (pp. 207-226). Brussels, Belgium: University of Brussels Press.

Spada, N., & Lightbown, P. (2008). Form-focused instruction: Isolated or integrated? *TESOL Quarterly*, *42*(2), 181-207. doi:10.1002/j.1545-7249.2008.tb00115.x

Spada, N., & Tomita, Y. (2010). Interactions between type of instruction and type of language feature: A meta-analysis. *Language Learning*, *60*(2), 263-308. doi:10.1111/j.1467-9922.2010.00562.x

Sparks, R., & Ganschow, L. (1995). A strong inference approach to causal factors in foreign language learning: A response to Maclntyre. *Modern Language Journal, 79*(2), 235-244. doi:10.1111/j.1540-4781.1995.tb05436.x

Suzuki, M. (2009). The compatibility of L2 learners' assessment of self- and peer revisions of writing with teachers' assessment. *TESOL Quarterly*, *43*(1), 137-148. doi:10.1002/j.1545-7249.2009.tb00233.x

Swain, M. (1985). Communicative competence: some roles of comprehensible input and comprehensible output in its development. In S. M. Gass & C. G. Madden (Eds.), *Input in second language acquisition* (pp. 235-253). Rowley, MA: Newbury House.

Swain, M. (1995). Three functions of output in second language learning. In G. Cook & B. Seidlhofer (Eds.), *Principle and practice in applied linguistics: Studies in honour of H. G. Widdowson* (pp. 125-144). Oxford, England: Oxford University Press.

Swain, M. (2005). The output hypothesis: theory and research. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 471-483). Mahwah, NJ: Erlbaum.

Swain, M. (2013). The inseparability of cognition and emotion in second language learning. *Language Teaching*, *46*(2), 195-207. doi:10.1017/S0261444811000486

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.

Taguchi, N. (2007). Task difficulty in oral speech act production. *Applied Linguistics*, *28*(1), 113-135. doi:10.1093/applin/am1051.

Taipjutorus, W., Hansen, S., & Brown, M. (2012). Investigating a relationship between learner control and self-efficacy in an online learning environment. *Journal of Open, Flexible and Distance Learning*, *16*(1), 56-69.

Takanashi, Y. (2004). TEFL and communication styles in Japanese culture. *Language, Culture and Curriculum*, *17*(1), 1-14. doi:10.1080/07908310408666678

Tavakoli, P. (2016). Fluency in monologic and dialogic task performance: challenges in defining and measuring L2 fluency. *International Review of Applied Linguistics in Language Teaching*, 54(2), 133-150. doi:10.1515/iral20169994

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. 239-276). Amsterdam, The Netherlands: Benjamins.

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

Thompson, G. (1996). Some misconceptions about communicative language teaching. *ELT Journal*, *50*(1), 9-15. doi.org/10.1093/elt/50.1.9

Towell, R. (2002). Relative degrees of fluency: A comparative case study of advanced learners of French. *International Review of Applied Linguistics*, *40*(2), 117-150. doi:10.1515/iral.2002.005

Towell, R., Hawkins, R., & Bazergui, N. (1996). The development of fluency in advanced learners of French. *Applied Linguistics*, *17*(1), 84-119. doi:10.1093/applin/17.1.84

Trang, T. T. T., Baldauf Jr., R. B., & Moni, K. (2013). Investigating the development of foreign language anxiety: An autobiographical approach. *Journal of Multilingual and Multicultural Development*, *34*(7), 709-726. doi:10.1080/01434632.2013.796959

Trofimovich, P., Lightbown, P. M., & Halter, R. (2013). Are certain types of instruction better for certain learners? *System*, *41*(4), 914-922. doi:10.1016/j.system.2013.09.004

Trofimovich, P., Isaacs, T., Kennedy, S., Saito, K., & Crowther, D. (2016). Flawed self-assessment: Investigating self- and other-perception of second language speech. *Bilingualism: Language and Cognition*, *19*(1), 122-140. doi:10.1017/S1366728914000832

Van Batenburg, E. S. L., Oostdam, R. J., Van Gelderen, A. J. S., Fukkink, R. G., & de Jong, N. H. (2019). Oral interaction in the EFL classroom: The effects of instructional focus and task type on learner affect. *The Modern Language Journal*, *103*(1), 308-326, doi:10.1111/modl.12545

Van Gelderen, A. (1994). Prediction of global ratings of fluency and delivery in narrative discourse by linguistic and phonetic measures-oral performances of students aged 11-12 years. *Language Testing*, *11*(3), 291-319. doi:10.1177/026553229401100304

VanPatten, B. (1993). Grammar teaching for the acquisition-rich classroom. *Foreign Language Annals*, *26*(4), 435-350. doi:10.1111/j.1944-9720.1993.tb01179.x

Vercellotti, M. (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

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Houndgrave, Hampshire, England: Palgrave MacMillan.

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

Wigglesworth, G. (2008). Task and performance based assessment. In E. Shohamy & N. H. Hornberger (Eds.), *Language testing and assessment: Encyclopedia of language and education, Volume 7* (pp. 111-122). New York, NY: Springer.

Wigglesworth, G., & Elder, C. (2010). An investigation of the effectiveness and validity of planning time in speaking test tasks. *Language Assessment Quarterly*, *7*(1), 1-24. doi:10.1080/15434300903031779

Williams, J. (2005). Form-focused instruction. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 671-691). New York, NY: Routledge.

Wolfe, E. W., & Smith, E. V. Jr. (2007). Instrument development tools and activities for measure validation using Rasch models: Part II-Validation activities. *Journal of Applied Measurement*, *8*(2), 204-234.

Wright, B. D. (1996). Time 1 to Time 2 (pre-test to post-test) comparison: Racking and stacking. *Rasch Measurement Transaction*, *10*(1), 478. Retrieved from https://www.rasch.org/rmt/rmt101f.htm

Wright, B. D. (2003). Rack and stack: Time 1 vs. Time 2 or pre-test vs. post-test. *Rasch Measurement Transactions, 17*(1), 905-906. Retrieved from https://www.rasch.org/rmt/rmt171a.htm

Xi, X., & Mollaun, P. (2006). Investigating the utility of analytic scoring for the TOEFL academic speaking test (TAST). *TOEFL iBT Research Report*. Retrieved from https://files.eric.ed.gov/fulltext/EJ1111433.pdf

Xu. H., & Lyster, R. (2014). Differential effects of explicit form-focused instruction on morphosyntactic development. *Language Awareness*, *23*(1-2), 106-121. doi:10.1080/09658416.2013.863899

Yan, J. X., & Horwitz, E. K. (2008). Learners' perceptions of how anxiety interacts with personal and instructional factors to influence their achievement in English: A qualitative analysis of EFL learners in China. *Language Learning*, *58*(1), 151-183. doi:10.1111/j.1467-9922.2007.00437.x

Yang, N. (1999). The relationship between EFL learners' beliefs and learning strategy use. *System*, *27*(4), 515-535. doi:10.1016/S0346-251X(99)00048-2

Yang, M., Badger, R., & Yu, Z. (2006). A comparative study of peer and teacher feedback in a Chinese EFL writing class. *Journal of Second Language Writing*, *15*(3), 179-200. doi:10.1016/j.jslw.2006.09.004

Yang, Y., & Lyster, R. (2010). Effects of form-focused practice and feedback on Chinese EFL learners' acquisition of regular and irregular past tense forms. *Studies in Second Language Acquisition*, *32*(2), 235-263. doi:10.1017/S272263109990519

Yashima, T. (2002). Willingness to communicate in a second language: The Japanese EFL context. *Modern Language Journal, 86*(1), 54-66. doi:10.1111/1540-4781.00136

Yashima, T., Zenuk-Nishide, L., & Shimizu, K. (2004). The influence of attitudes and affect on willingness to communicate and second language communication. *Language Learning, 54*(1), 119-152. doi:10.1111/j.1467-9922.2004.00250.x

Young, D. J. (1991). Creating a low-anxiety classroom environment: What does language anxiety research suggest? *The Modern Language Journal*, *75*(4), 426-439. doi:10.1111/j.1540-4781.1991.tb05378.x

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

# APPENDICES

**APPENDIX A**

**CONSENT FORM (JAPANESE VERSION)**

研究参加**同意書**
2015年4月

本状は英語スピーキング能力についての研究へのご協力をお願いするものです。研究目的は大学生の英語スピーキング能力の向上です。日本の英語教育を向上させるためには、皆様の協力が必要です。以下の条件をご理解の上、ご協力をお願い致します。

- 協力は任意です。
- 研究に関する質問はいつでもお受けします。
- 研究結果に関するすべての論文・発表において、本人を特定する情報が公表されることはありません。
- この研究に参加される内容や結果が授業成績に反映されることはありません。

私は Shzh-chen Nancy Lee の、speaking proficiency development についての研究に協力することに同意します。

署名_____

日付_____ (e.g. 2015/4/25)

英語スピーキング能力の発達をよく知るために、Interviewee も募集しています。
Interview に参加可能な方、<u>可能</u>に○をつけてください。可能

## APPENDIX B

## CONSENT FORM (ENGLISH VERSION)

Research Participation Consent Form
April 2015

You are being asked to participate in a research on the development of English speaking proficiency. The purpose of this research is to improve Japanese students' speaking ability. Your help is important in order to improve the current Japanese English education. Please understand the following points before participating in this research.

- Participation is voluntary.
- You are free to ask questions about this research anytime.
- Your personal information will not be disclosed in any paper or presentation related to this study.
- The results of your participation will not affect your course grade.

I agree to take part in the study on speaking proficiency development by Lee Shzh-chen Nancy

Signature_____Date……. (E.g. 2015/4/25)

I am looking for volunteers to participate in an interview to better understand English speaking development. Please circle Possible if you are able to participate in the interview.

Possible

## APPENDIX C

## BACKGROUND INFORMATION QUESTIONNAIRE

**✱1. Your number**

a

**✱2. Age**

s

**3. Gender**

- ○ Male
- ○ Female

**4. TOEFL**

**5. When did you take TOEFL (Year/Month) e.g. 2013/08**

**6. TOEIC**

**7. When did you take TOEIC (Year/Month)? e.g. 2013/08**

**8. IELTS**

**9. When did you take IELTS (Year/Month)? e.g. 2013/08**

**10. EIKEN**

**11. When did you take EIKEN (Year/Month)? e.g. 2013/08**

**12. What was your Hensachi? If you remember**

English studying experience

**13. Did you study English in elementary school? If yes, please describe your experience in English or Japanese**

**✱14. Describe your experience with English in junior high school in Japanese or English (e.g. conversation tasks, grammar translation tasks)**

d

**✱15. Describe your experience in high school in Japanese or English (e.g. conversation tasks, grammar translation tasks)**

d

**16. Were you (or are you) in clubs in which you use(d) English)? If yes, describe your experience in Japanese or English**

Prev Next

17. What are your overseas experience? Please write down 1) country, 2) purpose of visit, e.g. holiday or english study, 3) when did you go, 4) length of your stay e.g. 2 weeks

18. Have you studied English outside of school? E.g. Juku, Eikaiwa, Private English lessons. When did you study and for how long? E.g. Juku for about 3 years from 2007-2009.

*19. Please describe characteristics of a good English speaker. Please explain in Japanese or English.
(あなたは、英語を話すのが堪能な人とはどのような人だと思いますか？
堪能な人の特徴について、以下に詳しく説明して下さい)

## APPENDIX D

## THREE TARGET FORMS FOR DEDUCTIVE BLACKBOARD TEACHING

**Form 1:**
Example: A woman watched a TV program on eco tours

**Explanation**
Mozart was an Austrian musician and composer. He lived from 1756 to 1791. He started composing at the age of five and wrote more than 600 pieces of music. He was only 35 years old when he died.

**Lived**/**started**/**wrote**/**was**/**died** are all .

**Form 2: Past continuous**
Example: A woman was watching a TV program on eco tours

**Explanation**
Yesterday, Karen and Jim played tennis. They began at 10 o'clock and finished at 11:30. They were playing = they were in the middle of playing. They had not finished playing. **Was/were -ing** is the past continuous tense.

**Form 3: Past perfect**
Example: A woman had watched a TV program on eco tours

**Explanation**
Sarah went to a party last week. Paul went to the party too, but they didn't see each other. Paul left the party at 10:30 and Sarah arrived at 11 o'clock. So, when Sarah arrived at the party, Paul wasn't there anymore. He had gone home. **Had done** is past perfect tense.

I/we/they/you/she/he/it/Sarah + **had** + gone /seen / finished (**past participle**) etc.

## APPENDIX E

## GRAMMAR WORKSHEET

**(I did)**

Activity A: Read what Laura says about her work day:

| |
|---|
| I usually get up at 7 o'clock and have a big breakfast. I walk to work, which takes about 30 minutes. I start work at 9am. I never have lunch. I finish work at 5 o'clock. I am always tired when I get home. I usually cook a meal in the evening. I don't go to bed until 11 o'clock and I always sleep well. |

Yesterday was a typical working day for Laura. Write what she did or didn't do yesterday.

1. She _____ at 7'clock.
2. She _____ a big breakfast.
3. It _____ to get to work.
4. She _____ tired when _____ home.
5. She _____ a meal yesterday evening.
6. She _____ out yesterday evening.
7. She _____at 11 o'clock.
8. She _____ well last night.

Activity B: Complete the sentence using following verbs in correct form:

| |
|---|
| buy catch cost fall hurt sell spend teach throw write |

Mozart wrote more than 600 pieces of music.

1. How did you learn to drive? My father _____ me.
2. We couldn't afford to keep the car so we _____ it.
3. Dave _____ his legs.
4. Ann _____ a lot of money yesterday.
5. She _____ a dress which _____ $10.

Activity C: Complete the sentences by putting verb into the correct form, positive or negative.

1. It was warm so I _____ off my cost (take).
2. The film wasn't very good. I _____ it (enjoy).
3. The window was open and a bird _____ into the room (fly).
4. I was in a hurry so I _____ time to call you (have).
5. The bed was uncomfortable. I _____ well (sleep).

**Past continuous (I was doing)**

Activity A: What were you doing at these times? Write sentences as in the examples.

Example (at 8 o'clock yesterday evening) I was having dinner.

1. (at 5 o'clock last Monday afternoon) _____.
2. (at 10.15 yesterday morning) _____.
3. (at 4:3 this morning) _____.
4. (one hour ago) _____.
5. (last Saturday) _____.

Activity B: Complete the sentences in past continuous using your own ideas.
Example. Matt phoned while we were having dinner.

1. The doorbell rang while I _____.
2. We saw an accident while we _____.
3. She looked happy when she was _____.
4. The TV was on but nobody was _____.
5. The postman came when _____.

Activity C: Complete the sentences by putting verb into the correct or continuous forms.

Example. Jenny was waiting (wait) for me when I arrived (arrive)

1. What _____ (you/do) yesterday? I was asleep.
2. How fast _____ (you /drive) when the accident _____ (happen)?
3. We were in a difficult position. We _____ (not / know) what to do.
4. When I was young, I _____ (want) to be a police woman.
5. Last night, I _____ (drop) a cup when I was _____ (wash) dishes.

**Past perfect (I had done)**

Activity A: Complete sentences using words in brackets.
Example. (You / go out) You had done out.

1. (it / change / a lot) It _____.
2. (she / arrange / to do something else) She _____.
3. (the film/ already / start) The film _____.
4. (she / just / have / breakfast) She _____.
5. (I / not / see / him for many years) I_____.

Activity B: Use sentences on the left to complete paragraph on the right

| Original tense | Past perfect tense |
|---|---|
| Someone broke into the office during the night.<br>We arrived at work in the morning. We called the police. | We arrived at work in the morning and found that someone had broken into the office the night. So, we called the police. |
| Laura went out this morning.<br>I rang her doorbell.<br>There was no answer. | I went to Laura's house this morning and rang her doorbell. But there was no answer. She _____ out. |
| Jim came back from holidays a few days ago.<br>I met him at the same day.<br>He looked well. | I met Jim a few days ago. He _____ from holidays. He looked very well. |
| Kevin sent Sally lots of emails.<br>She never replied to him.<br>Yesterday he got a phone call from her.<br>He was very surprised. | Yesterday Kevin _____ from Sally. He was very surprised because he _____ lots of emails but she never replied to him. |

Activity B: Complete the sentences by putting verb into the correct (I did), or past perfect (I had done).

Example. Was Paul at the party when you arrived? No, he had gone (go) home.

1. I felt very tired when I got home, so I _____ (go) straight to bed.
2. The house very quiet when I got home. Everybody _____ (go) to bed.
3. Sorry I am late. The car _____ (break) down on the way.
4. I was late for class. The teacher _____ (finish) class.
5. I mixed up flight time so the flight _____ (depart) when I arrived.

## APPENDIX F

## PAIR VERBAL PRACTICE ON THREE TARGET FORMS

In pairs, change the verb into three past tense forms and make sentences with them.

Make the sentences verbally for your partners.

| Verb | Simple past | Past progressive | Past perfect |
|---|---|---|---|
| Give | | | |
| Hand over | | | |
| Receive | | | |
| Talk | | | |
| Fall | | | |
| Rescue | | | |
| Surprise | | | |
| Injure | | | |
| Remember | | | |
| Response | | | |

## APPENDIX G

## PEER FEEDBACK COMMENT SHEET (JAPANESE VERSION)

パートナーの録音を聴いてから具体的なコメントをしてください。できれば具体的にコメントしてください。

1. Fluency 流暢さ (e.g. スムーズ、とまらず、speedily、
繰り返されず、作り直さず、自然的かどうか)

| Comments |
|---|

2. Accuracy 正確さ (正確な文法、過去分詞 (I was going)、単純過去(I went)、過去完了(I had gone)、冠詞 (a, the)、 リズム、語彙、動詞と名詞の違い (e.g. product and produce)、word combination (e.g. strong heavy rain)

| Comments |
|---|

3. Complexity
難しさ(難しい文法、長い文章、幅広い語彙、抽象的よりも具体的)

| Comments |
|---|

## APPENDIX H

## PEER FEEDBACK COMMENT SHEET (ENGLISH VERSION)

Listen to your partner's recording and give specific comments. Try to make your comments as specific as possible.

1. Fluency (e.g. was it smooth, did it stop, was it speedily, did it have repeats or repairs, and was it natural or not).

| Comments |
| --- |
| |

2. Accuracy (e.g., accurate grammar, past continuous (I was going), simple past (I went), past perfect (I had gone), articles (a, the), rhythm, verb and noun usage difference (e.g. product and produce), and word combination (e.g. strong heavy rain).

| Comments |
| --- |
| |

3. Complexity (difficult grammar, long phrases, wide range of vocabulary, being specific over being general).

| Comments |
| --- |
| |

## APPENDIX I

## COMPLETED PEER FEEDBACK COMMENT SHEET

Peer review checklist list

**1. Fluency (e.g.スムーズ, とまらず, speedily, 繰り返されず, 作り直さず、自然的にかどうか)**

Comments
・ときどき止まる. ・ときどき言い直す. ・準備段階でもう少しsentenceを考えておきたい.
、m

**2. Accuracy (正確な文法, 過去分詞 (I was going)、単純過去形 (I went)、過去完了 (I had gone)、冠詞(a, the)発音、リズム、語彙、動詞と名詞の違い (e.g. product, produce), word combination (e.g. ~~strong~~ heavy rain)**

Comments
・過去形がときどき現在形になる. (過去分詞にしようとつとめるが故に. ミスをひき起こしている)

3. Complexity (難しい文法、長い文章、幅広い語彙、抽象的よりも具体的)

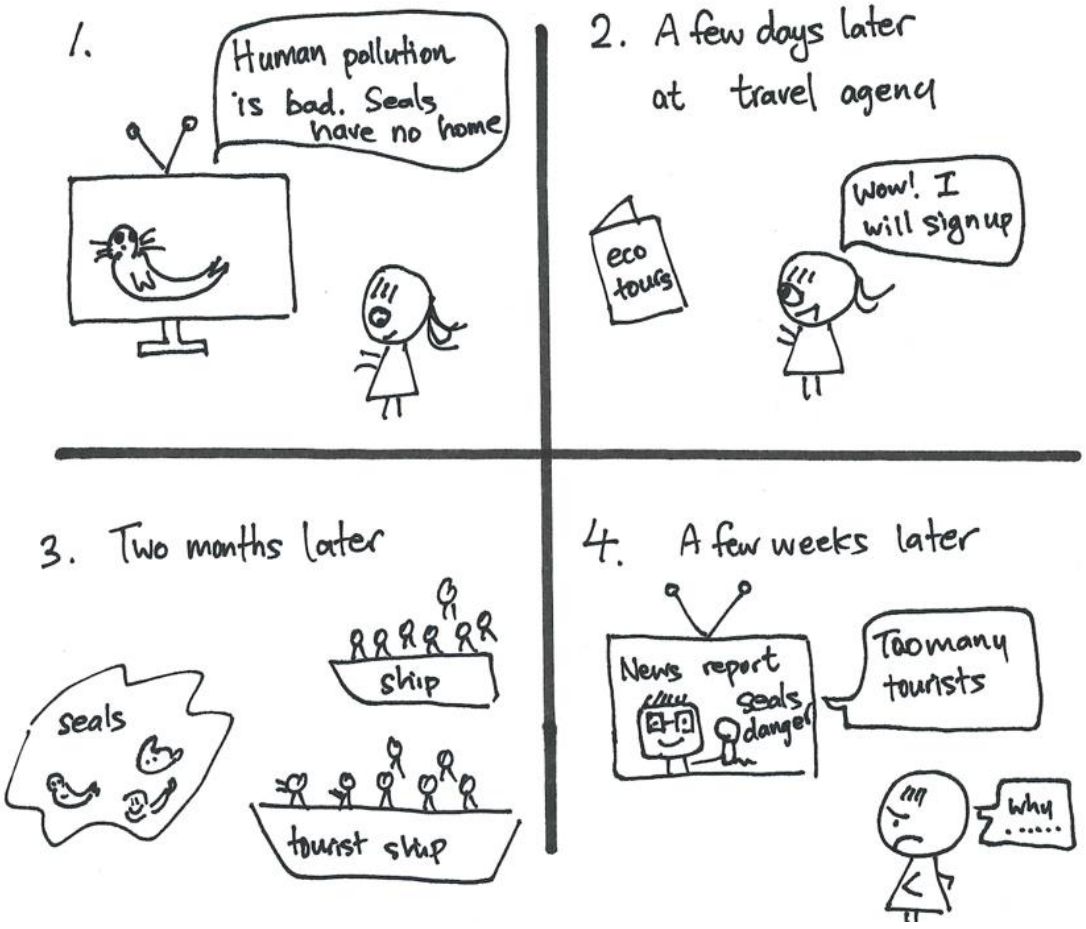
Comments
・工夫した言い回し (やや難しい単語を使用).

## APPENDIX J

## FOUR-PICTURE CARTOON FOR NARRATION

This is a story about a woman who participated in an eco-tour.
Your story should begin with the following sentence:

One day, a woman was watching a TV program.



## APPENDIX K

## SELF-ASSESSMENT OF SPEAKING TASK QUESTIONNAIRE

## (JAPANESE VERSION)

**Directions:**
あてはまる数字を選んでください。４コマの漫画ナレーションタスクを考え、答えて下さい。

| 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 |

**Fluency**

**1.** 私はスムーズに話すことができた。
**2.** 私はスムーズに話すことができた。
**3.** 私は少ないポーズで（間をあまり挟まずに）話すことができた。
**4.** 私はスピーディーに話すことができた。
**5.** 私は同じ単語を繰り返さずに話すことができた。
**6.** 私はどもらずに話すことができた。
**7.** 私は文章を作り直さずに話すことができた。
**8.** 私は自然に話すことができた。

**Syntactic Accuracy**

**9.** 私は正確な文法を用いて話すことができた。
**10.** 私は正確な動詞の時制を用いて話すことができた。
**11.** 私は正確な単純過去形を用いて話すことができた(例： I finished my homework)。
**12.** 私は正確な過去進行形を用いて話すことができた(例： I was doing my homework)
**13.** 私は正確な過去完了形を用いて話すことができた(例： I had finished my homework)。
**14.** 私は正確な冠詞を用いて話すことができた(例 a, the)。
**15.** 私は正確な動詞と名詞を用いて話すことができた。

**Syntactic Complexity**

**16.** 私は難しい文法を用いて話すことができた。
**17.** 私は長い文章を用いて話すことができた。
**18.** 私は色んな文章の組み合わせを用いて話すことができた。
**19.** 私は節を用いて話すことができた。

**20.** 私は慣れていない文章を用いて話すことができた。

## APPENDIX L

## SELF-ASSESSMENT OF SPEAKING TASK QUESTIONNAIRE

## (ENGLISH VERSION)

**Directions:** Please circle your answer from the choices below. Please refer to the four-picture cartoon when you answer the questions.

| 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| Strongly disagree | Disagree | Slightly disagree | Slightly agree | Agree | Strongly agree |

**Fluency**

1. I could speak smoothly.
2. I could speak with few pauses (without having much interval).
3. I could speak speedily.
4. I could speak without repetitions.
5. I could speak without stuttering.
6. I could speak without reconstructing sentences.
7. I could speak naturally.

**Syntactic Accuracy**

1. I could speak using accurate grammar.
2. I could speak using accurate verb tenses.
3. I could speak using accurate simple past tense (e.g., I finished my homework).
4. I could speak using accurate past continuous tense
(e.g., I was doing my homework).
5. I could speak using accurate past perfect tense
(e.g., I had finished my homework).
6. I could speak using accurate articles (e.g., a, the).
7. I could speak using accurate verbs and nouns.

**Syntactic Complexity**

1. I could speak using difficult grammar.
2. I could speak using long sentences.
3. I could speak using a combination of various sentences.
4. I could speak using clauses.
5. I could speak using unfamiliar sentences.

## APPENDIX M

## ANALYTICAL RATING SCALE FOR COMPLEXITY, ACCUACY, AND FLUENCY

| | Fluency |
|---|---|
| 5 | Speaks fairly fluently with only occasional hesitation, false starts and modification of attempted utterance. Speech is only slightly slower than that of a native speaker. |
| 4 | Speaks more slowly than a native speaker due to hesitations and word-finding delays. |
| 3 | A marked degree of hesitation due to word-finding delays or inability to phrase utterances easily. |
| 2 | Speech is quite disfluent due to frequent and lengthy hesitations or false starts. |
| 1 | Speech is so halting and fragmentary that conversation is impossible. |
| | **Accuracy** |
| 5 | Errors are not unusual, but rarely major. |
| 4 | Manages most common forms, with occasional errors, major errors present. |
| 3 | Limited linguistic control: major errors frequent. |
| 2 | Clear lack of linguistic control even of basic forms. |
| 1 | No linguistic control even of the most basic forms. |
| | **Complexity** |
| 5 | Attempts a variety of verb forms (e.g., passives, modals, tense, and aspect), even if the use is not always accurate. Takes risks grammatically in the service of expressing complex meaning. Regularly attempts the use of coordination and subordination to convey ideas that cannot be expressed in a single clause, even if the result is awkward or inaccurate. |
| 4 | Mostly relies on simple verb forms, with some attempt to use a greater variety of forms (e.g., passives, modals, more varied tense and aspect). Some attempt to use coordination and subordination to convey ideas that cannot be expressed in a single clause. |
| 3 | Produces numerous sentence fragments in a predictable set of simple clause structures. If coordination and/or subordination are attempted to express more complex clause relations, this is hesitant and done with difficulty. |
| 2 | Produces mostly sentence fragments and simple phrases. Little attempt to use any grammatical means to connect ideas across clauses. |
| 1 | No awareness of basic grammatical means. |

(Modified from Nitta & Nakatsuhara, 2014)

## APPENDIX N

## RATER GUIDELINES

### Purpose of Rating

Thank you for helping with the analytical rating section of my dissertation research. Your rating scores will only be used for my dissertation writing. The participants for rating are non-English major first-year Japanese students at a Japanese empirical university. Most participants have at least pre-intermediate English reading, listening and writing skills. Participants have a wide range in speaking proficiency, ranging from low to advanced proficiency levels. Please follow the following guideline.

### Pilot Rating

Before conducting rating, please complete the pilot rating using the attached audio files (a total of 15 one-min recordings). Please type scores into attached excel file and email it back to me. Please also email me if you have any questions about the rating criteria or process. Please proceed to the rating once you have completed the pilot rating and have no further questions.

### Instruction

1. Please make sure you can open all the audio files and sound quality are acceptable.
2. Rate participants' real performance and disregard bell curve rating.
3. Use the rating criteria (see Appendix B) when rating.
4. Type scores into the excel sheet prepared
5. Read the criteria before listening to each recording. Listen to the recording and spend no more than 30 seconds reading the rating criteria again before writing down a score for complexity, accuracy, and fluency respectively. Please do not spend too much time on each audio nor paying attention to the details. Just your gut intuitive judgement. You do not have to constantly look at the criteria sheet when listening to the recordings.
6. Do not listen to the recording twice. If possible, do not stop the recording half-way.
7. Each recording is approximately 90-120 seconds. There are about 100 recordings.
8. Please type comments or questions, if you have any into the excel sheet for each participant.

Please let me know if you have any questions.

## APPENDIX O

## ENGLISH SPEAKING LEARNER AFFECT QUESTIONNAIRE

## (JAPANESE VERSION)

**Directions:** あてはまる数字を選んでください。

| 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 |

**English Speaking Anxiety**

1. クラスの他の学生の方が自分より英語を話すのがうまいのではないかと心配になる。
2. クラス全体の前で英語を話すと不安になる。
3. グループで英語を話すと不安になる。
4. ペアで英語を話すと不安になる。
5. クラスメートと英語を話す時に間違いをするのではないかと心配になる。
6. 先生と英語を話す時に間違いをするのではないかと心配になる。
7. 先生から英語の質問が聞かれると不安になる。
8. 自分の意見を英語で表現できないと、あがってしまう。
9. 英語を話すと笑われるのではないか不安になる。
10. 英語で話されるとき、不安になる。
11. 先生が英語で何を話しているのか、分からないと不安になる。
12. 英語を話すのは、不安になる。
13. 日本語で質問をしない。
14. 先生の質問を英語で答えると恥ずかしくなる。
15. グループの中で、自分が英語を話す番が回ってくるとドキドキする。
16. グループ・ディスカッションで英語を話すと他の人が笑うのではないかと不安になる。

**English Speaking Self-efficacy**

1. レストランでの料理の注文を英語ですることができる。
2. 英語で自己紹介することができる。
3. 自分の趣味について英語で話すことができる。

4. グループの中で、英語で発表することができる。
5. クラス全体の前で、グループ発表を英語ですることができる。
6. クラス全体の前で、一人で発表を英語ですることができる。
7. 外国の人に英語で道案内を示すことができる。
8. 留学生に英語でキャンパス案内を示すことができる。
9. 英語で留学生と気軽に会話をすることができる。
10. 英語で先生と気軽に会話をすることができる。
11. 英語でクラスメートと気軽に会話をすることができる。

**Desire to Speak English**

1. 英語を話す国出身の友達を作りたい。
2. 留学生の友達を作りたい。
3. 色々な国から来た人と英語で話をしたい。
4. 日本人同士で英語で話をしたい。
5. 英語で自分の意見や考え方を表現したい。
6. 英語を話す国に行きたい。
7. 海外に行きたい。
8. 英語の授業でできるだけ英語を話したい。
9. 授業以外で英語を話す機会を見つけたい。
10. 留学生と英語で話をしたい。
11. 英語の先生と英語で話をしたい。

## APPENDIX P

## ENGLISH SPEAKING LEARNER AFFECT QUESTIONNAIRE

## (ENGLISH VERSION)

**Directions:** Please circle your answer from the choices below.

| 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| Strongly disagree | Disagree | Slightly disagree | Slightly agree | Agree | Strongly agree |

**Classroom English Speaking Anxiety**

1. I become worried when other students speak English better than I do.
2. I become anxious when I speak English in front of the entire class.
3. I become anxious when I speak English in small groups.
4. I become anxious when I speak English in pairs.
5. I become worried about making mistakes when talking in English with classmates.
6. I become worried about making mistakes when talking in English with the teacher.
7. I become anxious when the teacher asks me a question in English.
8. I become anxious when I cannot express my opinion in English.
9. I become anxious whether my classmates will laugh when I speak English
10. I become anxious when someone talks to me in English.
11. I become anxious when I don't understand what the teacher is talking in English.
12. I become anxious when I speak in English.
13. I do not ask questions in Japanese.
14. I become embarrassed when answering the teacher's question in English.
15. I can feel my heart pounding when it is my turn to speak English in the group.
16. I become anxious whether my classmates will laugh when I speak English in group discussions.

**English Speaking Self-Efficacy**

1. I can order food in English in a restaurant.
2. I can introduce myself in English.
3. I can talk about my hobbies in English.
4. I can give a presentation in English in groups.
5. I can give a group English presentation in front of the class.
6. I can give an individual English presentation in front of the class.
7. I can give street directions in English to foreigners.
8. I can give campus directions to international students.
9. I can talk with international students in English casually.
10. I can talk with the teacher in English casually.
11. I can talk with classmates in English casually.

**Desire to Speak English**

1. I want to make friends with people from English-speaking countries.
2. I want to make friends with international students.
3. I want to speak English with people from different countries.
4. I want to speak English with other Japanese people.
5. I want to express my opinions and ideas in English.
6. I want to visit English-speaking countries.
7. I want to go to overseas countries.
8. I want to speak English as much as possible in English class.
9. I want to find opportunities to speak English outside class.
10. I want to speak with international students in English.
11. I want to speak with the English teacher in English.

## APPENDIX Q

## NINE WEEKS OF NARRATION TRANSCRIPTS OF ONE PARTICIPANT

Week 1(Pretest)
One day a woman was watching a TV program. The name of woman is Lisa. Lisa saw saw the news that says human human pollution human pollution is destroying seals habitats. So so people people should pay attention to save them. And she she wanted to join the activity to to save the seal seal habitats seals habitats. A few days later, she found eco tours it helps helps save animals. Yeah and she she signed and she join the tour. Two two months later, she join the tour and she helped save the animals. But a few weeks later, she saw she saw news. It it said too many eco tourists were in danger and ship of tours sink so so she was very sorry to hear the news.

Week 2
One day, a mayor was meeting with other city hall officials|. They found that [eeee they they use too much electricity]| and and mmm they thought the [the cost of electricity is too too high].| So so they decrease decrease expenditure. Yes, maybe maybe in summer they they must use air conditioner| so elect electricity expenditure is likely to be high|. A few weeks later a few weeks later, eeeee they they decide that [they turn turn off all all lights during lunch hour]|. Eeee the next day at lunch time, They said [they turn turn off all lights at city hall]|. A month later, eeee they they succeed to cut cut down electricity consumption by five percent. |Maybe eee maybe if they cut down, no, maybe if they don't use eeee nandaro air conditioner[ they can eeee cut down more consumption]|.

Week 3
One day, an old lady was having a birthday party at her home. And then eee she was given she was gaved a cell phone as birthday party eee from her son. Eee She say she say she was very glad for it. And the next day, but the next day, she confess that it is too complicated for her to use use the cell phone. She even said I don't need a cell phone. Eeee two weeks later when eeee she went she went hiking eee with her friend and and eeee when she was enjoying the hiking, she fell down and and hurt her hurt her leg. A moment later, eee eeee her friend could call emergency service because she eee she has the cell phone. The friend didn't didn't't have any cellphone so if she didn't if she wasn't gave that cellphone at her birthday, they they could not eee they could have not call the emergency service. Eeee so the cellphone was very helpful at that time. EEee She she was glad glad to having the cellphone eee.

Week 4
One day, an old lady was having a birthday party at her home. And then eee she was given she was gaved a cell phone as birthday party eee from her son. Eee She say she say she was very glad for it. And the next day, but the next day, she confess that it is too complicated for her to use use the cell phone. She even said I don't need a cell phone.

Eeee two weeks later when eeee she went she went hiking eee with her friend and and eeee when she was enjoying the hiking, she fell down and and hurt her hurt her leg. A moment later, eee eeee her friend could call emergency service because she eee she has the cell phone. The friend didn't didn't have any cellphone so if she didn't if she wasn't gave that cellphone at her birthday, they they could not eee they could have not call the emergency service. Eeee so the cellphone was very helpful at that time. EEee She she was glad glad to having the cellphone eee.

Week 5

One day, a woman was talking to her parents. Eeee she confess that she was followed eee she was being mmmm followed chigau she was followed by someone when she was going going to her home. And and eee the parents were very surprised. And and they wanted to do something. And they decided to mmm consult a police. The next day, eee she asked a police to put a patrol on the streets at night. But but the police answer negatively. Eeee he he said that they didn't have enough po officers so so they want wanted eeee conduct the patrol but but they couldn't. A few weeks later, on resident meeting eee a man suggested that suggest to initiate eeee citizens patrol on on the street at night. And and everyone agreed with him. So a month later, the citizen citizen patrol was initiated and it made her relieved.

Week 6

One day a company employee a company employee came home late. Eee the time was almost midnight twelve o'clock. He had wanted eee he had wanted with his son or and daughter. Eeee he was very depressed. Eee yes his work was very tired. EEe was very busy and and there was a lot there was there were a lot of things to do. So he he should have eee over overtime working should have had overtime working. Then a few days later, he he confess he confess his manager that eee that he have has no time to spend with his family. And asked asked him to eee reduce the working time. A week a week later eeee in the company in the company meeting, the eee employer decided that eee they they reduce eee overtime working. Eeee The aim was eee improve working efficiency. And everyone was agree mmm agreed with it. So, a month later, he could he could go home by seven oclock and he could eat dinner with his family and talk with with them and it was enjoyable.

Week 7

One day a young was working in the head office of a large company. The company is very famous eee so so many people want to enter enter the company. But she was not satisfied with her job. Because it was not eee it was not what she really want to do. Eee one evening, her boss asked her eee to work overtime. But she she took down and she left work at five thirty. It was early. But but after that, she went to went to a dress making school. Yes. Eee it was it was making dress eeee it was making dress that what she she want to really. EEe she continue to learn to design eee dress for woman for one year mmm a year. EEe of course she of course she continued her job and she eee she

studied very hard and to put in effort. One year later, she went she became a eeee dress designer for women. Her dress is mmmmm famous and popular for many women.

Week 8
One day a boy was playing video games with friends in this room. And they were very they were very enjoying. But it it was very sunny. So eee his mother recommended them to eee go out and play and play soccer or soccer because eeee she thought they should exercise exercise. Half an hour later, eee they went to a park and near his house but there was a sign. It say it was saying, no ball games. So eeee they were not allowed to play soccer. But they started to they started to play soccer. And but a little later, it was eee it was very exciting to for them to play soccer. So they were were playing out of park. So so they bothered a driver in in the road kana on the road. So, eee the next day, they they stopped eee playing soccer in a park. Uuuu they added to play video games eee in the park.

Week 9 (Posttest)
One day a woman was watching a TV program. Eee when the program was saying that human pollution was destroying seals habitats. And she was interested in eeee improving enviro- environment because she likes likes world animals very much. A few days later, she went to a travel agency and found eee eco tours eee that have world animals eeee by cleaning environment. So eee she attr- she was attracted by the tours and she decided to join. Eeee so two months later, she really went to the tours and she did many activities to improve the environment and save the world animals. Eeee she was very she was enjoying and satisfied. But a few weeks later a few weeks later, when she was watching the TV program, the program said too many eco tours eeee was were in dangers to sea. She was disappointed.