INVESTIGATING FLUENCY AND INTERACTIONAL COMPETENCE DEVELOPMENT IN A JAPANESE UNIVERSITY EFL DISCUSSION COURSE

A Dissertation
Submitted to
the Temple University Graduate Board

In Partial Fulfillment
of the Requirements for the Degree
DOCTOR OF PHILOSOPHY

---

by
Paul Michael Garside
August 2025

Examining Committee Members:

Timothy Doe, Advisory Chair, Teaching and Learning
David Beglar, Teaching and Learning
Eton Churchill, Teaching and Learning
Edward Schaefer, External Reader, Ochanomizu Women's University

ABSTRACT

Many learners express the desire to speak a language fluently, although in practice this goal entails not only speaking relatively quickly and smoothly, but also interacting in ways that are appropriate to the communicative context. However, it remains an open question whether and how quickly learners can acquire such skills in an EFL environment, in which opportunities to use the target language are often limited outside the classroom. The main purpose of this study was to investigate the degree to which Japanese university students' speaking fluency and interactional competence (i.e., the ability to interact) developed over the course of one academic semester during a communicative, twice-weekly speaking and discussion course. Another purpose was to investigate any changes in speaking self-efficacy that occurred during the course, and to relate this variable to the fluency and interactional competence data. Two additional individual difference variables—speaking anxiety and extraversion—were also investigated. The participants were 20 low-intermediate learners, selected from two intact classes, with data gathered at three time points throughout the academic semester.

A mixed methods research design was used to investigate these issues. Monologue activities were used to measure utterance fluency, which represents the quantifiable features of fluency, such as speed and pausing. To triangulate fluency gains and to investigate performance in the interactive context, fluency measures were also taken from group discussions. These discussions were also used for native-speaker ratings of perceived fluency. A combination of quantitative and qualitative analysis was used to investigate interactional competence development, also based on performance in

the group discussions. Finally, questionnaire data were used to investigate the individual difference variables of speaking self-efficacy, speaking anxiety, and extraversion.

Results indicated that speaking fluency significantly improved over the course of the semester. Early gains in utterance fluency were mainly due to a reduction in pausing time. In the second half of the semester, however, an accumulation of marginal gains across various measures indicated that discrete aspects of speech production were integrating and thus working more efficiently. Furthermore, these utterance fluency gains were matched by gains made in the interactive context, especially in the first half of the course. Specifically, there was an increase in the number of syllables produced during the group discussions, as well as a reduction in the number and length of inter-turn pauses. These gains were also detected by the raters, who judged the participants significantly higher in fluency at the end of the course. Taken together, these findings present a powerful case for the fluency-enhancing benefits of communicative interaction in an EFL environment.

Results also indicated that interactional competence developed during the semester. Quantitative analysis revealed a significant increase in the number of turn and topic management moves over time, indicating an increase in collaborative behavior as well as a greater willingness to take the initiative in managing the interaction. In contrast, there was a significant decrease in the number of topic extension (i.e., content generating) statements, with the participants producing longer and more complex contributions over time as they adapted to the demands of the discussion tasks. These developments appeared to create a positive feedback loop, as longer topic extensions and more efficient turn-taking facilitated more speaking practice, thus creating the conditions necessary for fluency development.

The development of interactional competence was further highlighted by qualitative analysis, which confirmed the shift from short, often sub-clausal, topic extensions to longer, multi-clausal extensions. It also revealed a gradual progression towards the smoother use of turn and topic management features, and a diversification of techniques for marking agreement and disagreement. In addition, techniques for negotiating meaning were applied to a broader range of situations, and several instances of collaborative turn completions emerged at the final data collection point. Finally, there was a reduced reliance on L1 when negotiating meaning and transitioning between turns and topics. Many of these skills and techniques were presented and practiced during the course; moreover, their use was not related to the participants' levels of fluency, suggesting that these aspects of interactional competence can be directly taught.

The third main finding was that, although anxiety remained stable, speaking self-efficacy increased significantly in the first half of the course, indicating that affective benefits resulted as students grew accustomed to the communicative environment. This increased measure correlated with several of the performance variables—specifically utterance fluency, number of syllables produced in the discussions, and length of topic extension—at the mid point of the course. However, these positive correlations disappeared at the end of the course, indicating that students lower in positive affect had caught up with their peers in these aspects. Therefore, the results of this study indicated not only that fluency and interactional competence can develop during a single semester of communicative EFL instruction, but that the benefits of this pedagogical approach can extend to learners with various types of affective profiles. In terms of personality, however, extraverted students consistently performed more fluently, and participated more fully, than their introverted peers. Furthermore, the fact that the extraversion correlations were the most stable of all the individual difference variables over time supports the idea that, as a dispositional personality trait, its effects are more predictable than contextualized characteristics such as self-efficacy and anxiety.

ACKNOWLEDGMENTS

This study would not have been possible without the guidance and expert assistance of several people. First, I would like to thank my dissertation advisor, Dr. Tim Doe, who provided continuous support throughout this process. I have known Tim for many years, having initially worked together on a discussion-based course, and I was very fortunate to discover that we have a shared interest in the development of speaking fluency and interactional competence. We held many meetings during the course of this study, both in person and online, and I really appreciate the prompt and targeted feedback he provided at every stage.

I would also like to thank Dr. David Beglar, who has stimulated my interest in the field of second language development since I entered the Master's program at TUJ. I always found his lessons engaging and thought-provoking, and his deep knowledge of the field has been an inspiration. Furthermore, I appreciate the time he took out of his busy schedule to offer advice and feedback on my dissertation after I progressed to the doctoral program.

I am also grateful to the other members of my dissertation defense committee, especially Dr. Eton Churchill for his detailed feedback that went above and beyond what was required of his role. Both at the proposal and dissertation defense stages, he provided additional suggestions that improved the study in numerous ways. I also appreciate Dr. Edward Schaefer's kind encouragement and advice during the dissertation defense.

In addition, I would like to thank the other members of my doctoral cohort for their moral support and encouragement throughout this process. We shared many interesting discussions during the coursework phase of the program, which made attending lessons very enjoyable. Moreover, I feel grateful to have made some very good and knowledgeable friends, who stimulated my interest in a wide range of topics, extending beyond the field of language acquisition. Some of them also participated as raters, so I would like to extend particular thanks to them for contributing to the study directly. There are many other teachers, classmates, and administrative staff that I owe a debt of thanks to, all of whom have contributed to the unique community that TUJ provides.

Finally, I would like to give a special mention to my family. This study truly would not have been possible without their love and support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

In this chapter, I describe the background to the study and set out the main issue I aimed to investigate—the extent to which students in an English as a Foreign Language (EFL) context develop their speaking fluency, interactional competence, and speaking self-efficacy during a single semester of study in a communicative speaking class. After introducing the main constructs of interest, I briefly discuss the main theoretical and methodological issues involved in such research. I then outline the specific aims of this study, as well as its intended audience, delimitations, and the organization of the study.

Background of the Issue

One of the primary goals for many foreign language learners is the ability to speak fluently. Indeed, for many people spoken fluency is almost synonymous with language proficiency (Lennon, 1990). A simple definition of fluency is the ability to express one's ideas smoothly and quickly, although a more precise definition is necessary in order to quantify the construct. In other words, fluency has to be broken down into its various component parts before it can be measured and its development investigated.

Fluency is a multidimensional construct and in its broadest sense goes beyond linguistic features. It encompasses such wide-ranging elements as topic familiarity and cultural knowledge, because the ability to speak fluently presupposes adequate knowledge of appropriate content matter. Put simply, one cannot speak fluently, in any meaningful sense, without having something relevant to talk about. The multidimensional nature of fluency makes defining it a challenge for researchers, although Segalowitz's (2010) triadic framework, which divides the construct into cognitive, utterance, and perceived fluency, is widely used. Encompassing both cognitive and social perspectives, it provides a useful starting point and is referred to throughout this study.

Cognitive fluency refers to the efficiency with which a speaker integrates the various mental processes that underpin speech production (Segalowitz, 2010). Utterance fluency represents the observable, quantifiable features of fluency, such as speed and pausing, from which cognitive fluency can be inferred. In cognitive-based SLA research, utterance fluency is commonly subdivided into three further components: speed of delivery, which refers to how fast a sample of speech is articulated; breakdown fluency, represented by the number, duration, and placement of pauses; and repair fluency, which is the number of repetitions and self-corrections made by a speaker (Tavakoli & Skehan, 2005). Finally, perceived fluency refers to a listener's impression of a speaker's fluency. This element, although external to the learner, cannot be overlooked given the interactive nature of speech in most situations. In other words, speaking is an inherently social activity, and this characteristic should be reflected in any comprehensive definition of the construct. Furthermore, perceived fluency addresses the limitations of utterance fluency, as objective measurements do not capture whether speech appears fluent to a listener, which is likely to be influenced by the unquantified elements of pronunciation and prosody. To sum up, fluency encompasses all aspects of this triadic framework and therefore clearly involves more than just speaking quickly.

Given its status among learners, L2 fluency research has attracted growing interest in recent decades. Most of this research has been focused on individual monologue activities, despite the fact that speaking almost always takes place in an interactive context, with two or more people contributing to the discourse. There are obvious methodological advantages to this approach, as extended stretches of speech can be produced without interference from an interlocutor (Tavakoli & Wright, 2020). Any form of response or feedback can affect the output of a speaker, for instance causing them to pause or reformulate a comment if the listener indicates a lack of comprehension. Even non-verbal feedback, such as facial expressions, can influence a speaker's output, thereby undermining any inferences that can be drawn regarding the speaker's underlying cognitive fluency. Moreover, it cannot be assumed that research findings from one context transfer to another, so fluent performance during a monologue might not be repeated during interaction. There is, therefore, a strong case for measuring fluency not only during monologues but also in the kinds of social environments in which communicative discourse naturally takes place. However, such research has been lacking thus far.

One aspect of learner performance that is self-evidently related to communication is interactional competence (Kramsch, 1986), which focuses on turn-taking and problems of understanding (Wong & Waring, 2021), and emphasizes the social context of discursive practices (e.g., Young, 2011). Interactional competence has been defined as the ability "to participate in courses of action involving two or more speakers, and which crucially has to do with turn-taking dynamics" (Pallotti, 2019, pp. 6–7). This construct is inherently social, as any form of interaction is by definition co-constructed by its participants. It includes interactional resources, such as the ability to manage multi-party sequences, and has become increasingly recognized as an important aspect of second language development (Galaczi & Taylor, 2020). In theory, utilizing these resources should allow the group as a whole to construct more fluent interactions, with less pausing between turns, while maintaining mutual comprehension. According to Young (2011), "[Interactional competence] is how those resources are employed mutually and reciprocally by all participants in a particular discursive practice" (p. 428). In order to participate fully and actively in extended stretches of discourse, learners need to be aware of such techniques as how to initiate discourse, encourage others to participate, and overcome breakdowns in communication. These techniques can be broken down into various micro-features, such as asking questions, indicating agreement, and confirming comprehension, which can then be used for pedagogical or assessment purposes (Galaczi & Taylor, 2018).

Some foundational aspects of interaction, such as the cooperative system underpinning turn-taking, appear to be universal and hence transferrable from the L1 culture, with language-specific resources mapping onto these basic structures during L2 learning (Hall, 2018). However, the most salient point from a pedagogical perspective is that these features do not always emerge naturally in a second language. Interactional strategies and micro-skills, such as how to manage repair, can be directly taught (Richards, 1990), which enables learners to generate more cohesive output, as it provides them with a means of initiating and sustaining interaction. In other words, practicing and raising awareness of techniques related to interactional competence can build upon pre-existing competencies to improve the smoothness and efficiency of L2 interaction (Wong & Waring, 2021). Moreover, the existence of such culturally universal competencies means that low linguistic proficiency does not automatically imply low interactional competence (Waring, 2019). Even if the phrases used are basic and the delivery lacks smoothness, teaching learners to understand and use the elements that occur in effective interaction—such as indicating agreement or asking questions—allows even lower-proficiency learners to improve their ability to interact in the L2, thereby circumventing issues of linguistic proficiency.

These interactional features can also be quantified, although this point has proven controversial among researchers. For example, Schegloff (1993) argued that the contingent nature of interaction makes quantification of such features reductive, as whether a feature is produced or not depends not only on the ability or willingness of the speaker to produce it, but also on its relevance in that particular context. Schegloff's research was conducted with L1 data, however, and was therefore not concerned with learner development. Other researchers—working with L2 data—have maintained that highlighting how often different types of interactional moves (e.g., topic extensions) occur can illuminate the development of interactional competence, assuming that more competent speakers exhibit different patterns of interaction from their less competent peers. For instance, Galaczi (2008) found quantitative differences in the use of interactional features among learners of different proficiency levels.

Nevertheless, any quantification of such features has to be complemented by qualitative analysis. Quantification provides measurable information about such features, including how often they are used, but it does not show how effectively they have been used because an increase in use does not necessarily imply improvement (Lam, 2018). On the contrary, the overuse of an interactional strategy or formulaic phrase can be counterproductive and lead to "functionally impoverished" output (Lennon, 2000, p. 28), which, even if superficially fluent, is inappropriate for the demands of the task. Carefully selected exemplars of learner output can not only illustrate how L2 interaction proceeds but also whether particular phrases and techniques have been used effectively. In other words, these exemplars can be used to highlight and analyze the practices, such as turn-taking and repair, that represent the "observable incarnation" of interactional competence (Waring, 2018, p. 59).

As well as the linguistic variables of fluency and interactional competence, the individual difference variable of self-efficacy is also relevant to learner performance during interaction. Self-efficacy refers to the belief of individuals in their ability to successfully undertake an activity (Bandura, 1997). Given that fluency development is a goal for many learners, it is logical to assume that taking steps towards that goal would increase feelings of competence and self-efficacy. A positive feedback loop can result, as self-efficacy affects people's disposition, behavior, and motivation to continue an activity (Bandura, 1997). In the field of education, positive learning experiences are widely believed to have long-term motivational benefits (e.g., Dörnyei & Ushioda, 2011), so it is important to explore whether speaking self-efficacy develops during a course of study and, if so, how it relates to fluency and interactional competence.

Finally, regarding personality traits, it is often assumed that extraverts are better equipped to develop their L2 speaking ability than more introverted personality types. This belief is commonly held despite findings that extraversion is negatively related to academic success in general, as extraverts tend to develop less conscientious study habits than their introvert peers (O'Connor & Paunonen, 2007). However, the results of research investigating L2 speaking and extraversion are complex. Extraverts have been found to perform more fluently than introverts during speaking activities, especially in formal or stressful situations (Dewaele & Furnham, 1999). Consequently, L2 spoken interaction, with its potential for heightened stress and anxiety, is likely to place a higher burden on introverts. Therefore, whether the personality trait of extraversion is positively related to the performance variables of fluency and interactional competence is also important to ascertain.

Statement of the Problem

There is a growing body of research on the development of fluency as measured during monologic activities (e.g., De Jong & Perfetti, 2011; Nation, 1989). Researchers have looked at the effects of task repetition on fluency within the theoretical framework of skill acquisition theory, which focuses on the proceduralization and automatization of linguistic knowledge (Anderson, 1992; DeKeyser, 2007). This approach assumes the primacy of individual cognition in language acquisition and acknowledges that automaticity, and the retrieval of ready-made chunks of language, is an essential element of fluency. Nevertheless, speech is almost always produced in a social context; therefore, if the desired outcome is fluent performance during interaction, it is necessary to examine performance in such an environment, not only in the more controlled setting of monologue delivery.

Fluency is usually considered to be an individual attribute, although fluent interaction must be jointly constructed during multi-party talk as participants alternate between speaker and listener roles (Kasper, 2006). So integral is turn-taking to the smoothness and efficiency of interaction that McCarthy (2010) has coined the term *confluence*, described as “the joint production of flow by more than one speaker, focusing in particular on turn openings and closings” (p. 1). In other words, rather than producing a series of individual turns, successful interaction requires speakers to contribute to the flow of the discourse by responding contingently and relating their utterance to a previous speaker's turn. This conception of fluency as co-constructed during interaction underlines the need to examine it in social, as well as individual, contexts. Despite the appeals of numerous researchers (e.g., Tavakoli & Wright, 2020), such research has not been forthcoming, perhaps due to the methodological complications mentioned previously.

Interactional competence, on the other hand, can only be examined in social settings because it is only manifested during interaction (Roever & Kasper, 2018). Attempts to quantify it have tended to focus on high-stakes oral examinations, with coding schemes tailored to the tasks and kinds of language produced in such contexts (e.g., Galaczi, 2014). However, it has not been quantified during unassessed speaking activities in EFL classrooms. It is also unclear how interactional competence develops over time in the classroom, as studies based on high-stakes exams are by nature cross-sectional rather than longitudinal.

To what extent interactional competence pertains to the individual, or alternatively is reconstituted during each interaction, has also proven controversial (Bachman, 2007). This issue is of fundamental importance because it addresses the question of whether the knowledge acquired in one context can be redeployed in another. Some researchers have argued that interactional competence is entirely co-constructed and context-dependent, and hence unavailable for transfer to other situations (He & Young, 1998). Others have maintained that, although contextual constraints clearly exist, individuals bring their own resources—linguistic or otherwise—to any instance of communication, and that it is these resources that constitute interactional competence

(Fulcher, 2010). Given the need for learners to transfer their skills to different contexts, with different interlocutors, this latter view better informs the current study.

Pedagogical interventions can enhance communicative performance if learners are made aware of specific techniques, such as initiating repair, that enable them to sustain cohesive L2 discourse (Wong & Waring, 2021). Talking for extended periods, with fewer and shorter gaps between turns, represents more fluent output in the group context, whereas the increased use of interactional strategies demonstrates improved interactional competence. There is thus scope for investigating when and how such techniques are used by learners and how their use changes during a semester of study.

Self-efficacy is associated with positive educational outcomes, as students with higher self-efficacy tend to expend greater effort, persist in the face of setbacks, and have lower levels of anxiety than students with low self-efficacy (Pajares & Schunk, 2001). Bandura (1997) proposed four sources of self-efficacy—mastery experiences, vicarious experiences, verbal persuasion, and physiological and affective states—although he identified mastery experiences as the most powerful of the four because they provide the most direct evidence of the ability to succeed. L2 speaking self-efficacy can be fostered in classroom settings by giving learners repeated opportunities to exchange information and opinions with their peers in a comfortable environment (Mills, 2014). Given the perceived lack of motivation for studying English in Japanese universities (Ryan, 2009), and the link between self-efficacy and motivation (Bandura, 1997), a communicative speaking course that results in increased speaking self-efficacy could be beneficial for students in terms of motivation and ultimately speaking ability.

There is also scope to clarify the role of more stable personality traits in this kind of communicative language learning. Although research regarding the relationship between extraversion and language learning outcomes has been mixed, Dewaele and Furnham (1999) maintained that this ambiguity is due to researchers failing to account for the effect of context, such as not distinguishing between written and verbal language criteria. To clarify this picture, it is necessary to link the variable of interest to specific contexts and activities. For example, there is evidence that extraverts are more suited to open-ended group speaking activities (Nakatsuhara, 2013), such as the ones used in this study.

Purposes and Significance of the Study

The main purpose of this study is to investigate the degree to which Japanese EFL students' fluency and interactional competence develop during a one-semester communicative course. These two variables are investigated together because both could plausibly develop in such a course. Regarding fluency, extensive opportunities for speaking practice create the conditions necessary for development. In contrast, prior to university many students experience heavily form-focused English lessons with limited opportunities to speak (Koizumi et al., 2022; Nishino & Watanabe, 2008). The emphasis in such lessons is on reading and grammar, as they are primarily geared towards passing university entrance exams. Although this goal is important as a proximal objective, it differs from any ultimate goal of achieving spoken fluency.

Regarding interactional competence, development could be facilitated by the teaching of individual strategies and techniques. Interactional competence in this academic EFL context is operationalized as consisting of *topic development* and *support* moves. Topic development moves consist of statements and questions that drive the topic forward, as well as turn taking moves that facilitate transitions between speakers and topics. Support moves consist of backchannels, agree/disagree markers, confirmations of comprehension, and repairs of misunderstandings. The choice of interactional features to be analyzed reflects the roles of speaker and listener in this context, in which students are expected to express their own opinions, as well as respond to the opinions of others.

Examining how learner-language develops is a fundamental aspect of SLA, although measuring development during communicative courses has been under-researched to date. The major significance of this study lies in attempting to address this gap. Specifically, measuring the development of fluency and interactional competence, and how they relate to each other, allows the communicative methodology underpinning such courses to be analyzed. By extension, it also allows the effectiveness of activities included in communicative courses to be investigated.

An additional purpose is to explore the extent to which objective measures of fluency and interactional competence correlate with subjective ratings of fluency and communicative adequacy. Speech rate, hesitation, and repair phenomena are objective in that they can be measured and do not rely on qualitative judgement. Aspects of interactional competence, such as the types of contributions (e.g., topic extension statements and active listening responses) a participant makes, can also be quantified. The issue of whether such measures correspond with the impression made on listeners is examined by comparing them with the subjective judgements of raters. Positive relationships would suggest that the relevant features contribute to the impression made on listeners, thereby further supporting the pedagogical value of fluency development and/or the teaching of interactional techniques.

Another purpose of this study is to explore any changes in speaking self-efficacy and anxiety that occur during the course, as self-efficacy can trigger a virtuous cycle of positive feelings, motivation, and behavior, as well as reduced anxiety (Pajares & Schunk, 2001). Communicating directly in an L2 is known to be a face-threatening activity, which is likely to cause anxiety among students unfamiliar with predominantly English-speaking environments (Horwitz et al., 1986). Therefore, if improvements in speaking self-efficacy—and a reduction in anxiety—can be observed during a single semester of study, these findings would provide further support for the effectiveness of such a communicative course. The issue of whether the individual difference variables of speaking self-efficacy and anxiety, and the personality trait of extraversion, correlate with the performance variables of fluency and interactional competence is also examined.

A final purpose of the study is to explore the influence of the pedagogical intervention provided in this communicative, discussion-based course. Interactional skills and strategies were presented and practiced in pairs and small groups, with the aim of developing fluency and the ability to interact in the L2. Students were encouraged to use specific phrases and techniques, for example when giving opinions and asking follow-up questions. Analyzing how and when these phrases and strategies are used during unassessed interaction, as opposed to during a test, allows researchers to infer which ones are valued by the participants themselves. The issue of whether their use demonstrates improved interactional competence and/or contributes to more fluent output in the group context is also investigated. Presenting detailed exemplars of learner output from the beginning, middle, and end of the course helps to illuminate this process.

If positive outcomes are associated with the intervention implemented in this study, this finding would have important pedagogical implications. The explicit presentation and practice of strategies and techniques aimed at fostering L2 interactional competence provides a systematic way of focusing communicative classes beyond unstructured talking for its own sake (Dörnyei & Thurrell, 1994). For example, in this course students are taught interactional techniques such as how to give and support opinions, express agreement or disagreement, and deal with communication breakdowns. If these and other techniques appear to promote the development of speaking skills, this finding would have significance for curriculum designers and teachers alike.

Research Questions and Methodology

To address these issues, this one-semester study investigated the development of speaking fluency—as measured in monologue activities, group discussions, and as perceived by native-speaker raters—and interactional competence, as well as the relationship between these two variables. It also investigated the extent of any changes in speaking self-efficacy and anxiety. In addition, the relationships between speaking fluency and the individual difference variables of speaking self-efficacy, anxiety, and extraversion were examined. The relationships between these individual difference variables and the quantified aspects of interactional competence were also investigated.

The data for this study was collected from a twice-weekly speaking course for low-intermediate freshman university students. One lesson each week was discussion-based, in which interactional skills and techniques were introduced and practiced. The other lesson was textbook-based, although many opportunities for spoken communication, whether in pairs or small groups, were provided in both lessons. A repeated monologue activity was also used in both lessons.

Twenty participants were selected for analysis from two intact classes, with data gathered at three points in the semester—near the beginning, middle, and end. This longitudinal design therefore allowed the development of fluency, interactional competence, and self-efficacy to be tracked throughout the semester. The relationships between these variables were also examined to ascertain whether having a higher degree of one construct provided any advantage regarding another.

Data collection activities consisted of audio-recorded monologues and video-recorded group discussions, consisting of three members, which were used to investigate the development of fluency and interactional competence. All recorded monologues and discussions were transcribed, with the latter additionally coded in terms of topic development and support moves. The individual difference variables were investigated using self-reported questionnaire data.

Theoretical Perspective

The main theoretical perspective adopted in this study is that the performance variables of fluency and interactional competence can be quantified and that they represent distinct elements of speaking proficiency. Although fluency is often used interchangeably with general speaking ability (Lennon, 1990), many SLA researchers view it as part of the wider complexity, accuracy, and fluency (CAF) framework used for the comprehensive analysis of speech production. For measurement purposes, fluency is commonly divided into the three subcomponents of speed of delivery, breakdown fluency, and repair fluency (Tavakoli & Skehan, 2005). These elements can be tallied and fluency development inferred from the results. Various types of interactional moves, such as topic extensions and listener responses, can also be tallied during interaction (Galaczi, 2014). Patterns of use can then be observed and the development of interactional competence similarly inferred. Qualitative analysis complements the quantitative data by revealing how the interactional moves have been interpreted by other participants, and whether they have been used in pragmatically appropriate ways.

Audience for the Study

The findings of this study should be of interest to second language acquisition researchers, curriculum developers, test-designers, language teachers, and, by extension, second language learners. If the performance variables of fluency and interactional competence are influenced by pedagogical interventions, language practitioners can use this information to design courses that help learners develop these elements of their speaking ability. For example, interactional strategies and techniques can be integrated into lessons and form the basis of interactive speaking tests that incorporate such features in the rubric. Furthermore, if gains can be observed in a single semester, this finding should interest researchers concerned with observing how language develops in classroom contexts. Finally, if any gains are related to improvements in speaking self-efficacy and/or a reduction in anxiety, such findings should interest practitioners and learners interested in the affective dimension of language learning.

Delimitations

Delimitations refer to the external validity of a study; in other words, the degree to which it is generalizable or transferrable to other contexts. I hope this study benefits the audiences mentioned above and that it advances the understanding of how L2 speaking fluency and interactional competence develops in a communicative classroom environment, and how this development is related to the individual difference variables of interest. Inevitably, however, its generalizability depends on the context under consideration and it has greater relevance in similar contexts. Below, I outline the context of the current study and the implications for its generalizability.

First, all participants were drawn from one specific context, as they were all first-year students at a private Japanese university. Moreover, they were all Japanese students, with a shared linguistic and cultural background, which restricts the degree to which the results can be applied to other contexts. There is a considerable degree of linguistic distance between Japanese and English, which have very different phonological, lexical, syntactic, and morphological structures; therefore, there is no guarantee that speakers of other languages or cultural backgrounds would produce similar results. Furthermore, all students were drawn from the same proficiency level, as measured on the university's placement test (see Chapter 3: Participants), so the results cannot automatically be generalized to learners of other proficiency levels.

A further delimitation is in the use of students from intact classes rather than randomly assigned participants. A fully experimental design, with learners drawn at random from the population of interest, would be preferable in terms of enhancing generalizability. However, the fact that this study was conducted in a classroom setting adds a pedagogical authenticity that randomized, laboratory-based experiments sometimes lack in the field of educational research (DeKeyser & Botana, 2019). In other words, the fact that the participants were acquainted with each other reflects the reality of the classroom, as well as discourse in most naturalistic settings. Nevertheless, if group members had been unacquainted, different results could have been produced. For example, if speakers feel more relaxed in each other's company, gaps between turns could be shorter and they might feel more comfortable with repairing each other's utterances.

The fact that groups of three were used to gather the interaction-based data is another delimitation. This number was chosen as the intervention consists of a classroom-based discussion course and many of the techniques and phrases taught are less applicable to dyadic interaction. The interactional competence coding system in the current study (see Chapter 3) has been adapted from one specifically developed for high-stakes pair interaction (Galaczi, 2014) due to the differences between the two contexts. The fact that it has been modified for use in the current context means that the results should be interpreted cautiously, as the differences inherent in small group interaction mean that the results might not be applicable to dyadic or larger group contexts.

Organization of the Study

Chapter 2, Review of the Literature, consists of the following main sections: Defining Speaking Fluency, Researching L2 Speaking Fluency, Interactional Competence, Speaking Self-Efficacy, Language Anxiety, and the Role of Extraversion. At the end of the chapter, I present the gaps in the literature, purposes of the study, and research questions. In Chapter 3, Methods, I describe the context of the study, the participants, instrumentation, procedures, and analyses used. In Chapter 4, Preliminary Analyses, I describe how validity evidence was gathered and analyzed for the instruments. In Chapter 5, Results, I present the results of the study, which are interpreted in Chapter 6, Discussion. At the end of that chapter, I discuss the theoretical and pedagogical implications of the findings. Finally, in Chapter 7, Conclusion, I summarize the main findings of the study, point out its limitations, and offer suggestions for future research.

CHAPTER 2

REVIEW OF THE LITERATURE

In this chapter, I review the literature on fluency, interactional competence, and the individual difference variables of speaking self-efficacy, anxiety, and extraversion. At the end of the chapter, I identify gaps in the literature, state the purposes of the study, and list the research questions the study is designed to answer.

Defining Speaking Fluency

The multidimensional nature of fluency has made defining it a challenge for researchers. A simple definition is the ability to express one's ideas smoothly and quickly, although producing fluent speech is far from a simple achievement, as it requires sociolinguistic and pragmatic knowledge, in addition to sufficient knowledge of grammar and vocabulary (Koponen & Riggenbach, 2000). Although formulated for L1 research, Fillmore's (1979) wide-ranging definition of fluency is regarded as a seminal starting point:

> fluency is the ability to talk at length with few pauses, the ability to fill time with talk... the ability to talk in coherent and semantically dense sentences... the ability to have appropriate things to say in a wide range of contexts... [and] the ability some people have to be creative and imaginative in their language use (p. 51).

Another landmark definition directly related to L2 research was provided by Lennon (1990), who distinguished between broad and narrow senses of fluency. Broad fluency refers to general oral proficiency, often perceived holistically, whereas narrow fluency applies specifically to the flow of speech and hence the more objective and measurable elements. This distinction was expanded upon by Tavakoli and Hunter (2018), who used second-language teachers' definitions of fluency to construct a four-point model consisting of very broad (general L2 proficiency), broad (general L2 speaking ability), narrow (ease, flow, and continuity—as distinct from grammatical complexity and accuracy), and very narrow (the measurable aspects of fluency).

Among SLA researchers, Segalowitz's (2010) triadic framework is widely used because it encompasses both cognitive and social aspects of fluency. This framework consists of three intertwined elements: cognitive fluency, utterance fluency, and perceived fluency. Taking cognitive fluency first, this element refers to the efficiency with which speakers integrate the various mental processes that underpin speech production. According to Suzuki and Kormos (2023), cognitive fluency consists of two dimensions: speed of processing and linguistic resources. Sentence construction (i.e., syntactic encoding) is the key to the former, whereas productive vocabulary size is the main component of the latter, as efficient output also depends on the breadth and depth of semantic knowledge. Due to the difficulty in directly measuring cognitive processes, little research has been conducted with relation to L2 speaking; however, some researchers have used reaction times to written or visual prompts to investigate how efficiently individuals process language. These measures can then be compared with utterance fluency measures (see below) to ascertain whether underlying cognitive processes are related to the observable aspects of speech production (e.g., Kahng, 2024; Segalowitz & Freed, 2004).

Utterance fluency represents the observable, quantifiable features of fluency in its narrowest sense. Cognitive fluency can then be inferred from temporal measures, such as speed and pausing, as more efficient processing should lead to more efficient output. In cognitive-based SLA research, utterance fluency is commonly subdivided into three further components: speed of delivery, which refers to how fast a sample of speech is articulated; breakdown fluency, represented by the number, duration, and placement of pauses; and repair fluency, which is the number of repetitions and self-corrections made by a speaker (Tavakoli & Skehan, 2005). However, the wide range of possible variables can make generalizing findings difficult, as researchers have not been consistent in the variables chosen or the measures used, such as minimum pause duration (De Jong & Bosker, 2013).

Finally, perceived fluency refers to a listener's subjective impression of a speaker's fluency. In other words, it is the inference that listeners make about a speaker's cognitive fluency based on the utterance fluency of a speech sample (Segalowitz, 2010), thereby encompassing all aspects of the triadic framework. Its importance lies in highlighting the social dimension of speech and the fact that fluency is more than a learner-internal, psycholinguistic construct (Tavakoli & Wright, 2020). Measures of perceived fluency can be obtained by having raters use a rubric to assess a speaker's fluency. Subjective ratings have been found to correlate with objective pause and speed measures of fluency (Bosker et al., 2013), as well as with speakers' self-ratings (Prefontaine, 2013). However, research has also shown the difficulty of obtaining reliable measures of perceived fluency as raters sometimes prioritize different factors, such as speech rate, lexical choice, and accuracy, when making their judgements (Kormos & Denes, 2004). Such disparities highlight the importance of providing training to ensure that raters are assessing the same construct.

Speech Production

As mentioned above, cognitive fluency refers to the efficiency with which speakers integrate the mental processes that underpin speech production. These processes have been divided into distinct stages, most influentially in Levelt's (1989, 1999) speaking model (see Figure 1). Although originally devised for L1 speech, it has been adapted and applied to L2 speech production (De Bot, 1992; Kormos, 2006), dividing the process into three stages: conceptualization, formulation, and articulation. In Levelt's model, these three stages are modular; that is, they are hypothesized to operate relatively autonomously, sequentially transmitting their output to the next stage of production. In proficient speech these processes operate in parallel, almost simultaneously, and with little cognitive effort. This system explains the ability of humans to talk smoothly and rapidly in languages in which they are proficient. On the other hand, in less developed—particularly L2—speech, the processes are not yet automatic and are more likely to operate serially, requiring greater time and effort (Tavakoli & Wright, 2020).

The first stage of the process is conceptualization of the message, in other words deciding what to say based on the speaker's knowledge of the world as well as knowledge of the interlocutor's state of mind. This stage is language-independent because it involves generating concepts rather than linguistic information; therefore, no L2-specific fluency issues occur at this stage (Segalowitz, 2010). According to Levelt's (1989) model, this pre-verbal message is then passed on to the next stage, the formulator, where it is encoded into grammatical, lexical, and phonological form. It is here that vocabulary knowledge is required and rules are applied, such as about verb tense or plural marking on nouns, which has obvious implications for L2 fluency. Finally, this speech plan progresses to the articulator, where it is converted into overt speech, requiring motor control of the muscles and physical features that produce speech sounds. This stage can also affect L2 fluency if these sounds cannot be articulated smoothly or with appropriate pronunciation. At every stage of this process, the various types of output are monitored and mistakes or unintended forms can be repaired or adjusted as necessary. This monitoring process also has implications for the rate of speech that is ultimately articulated and produced.

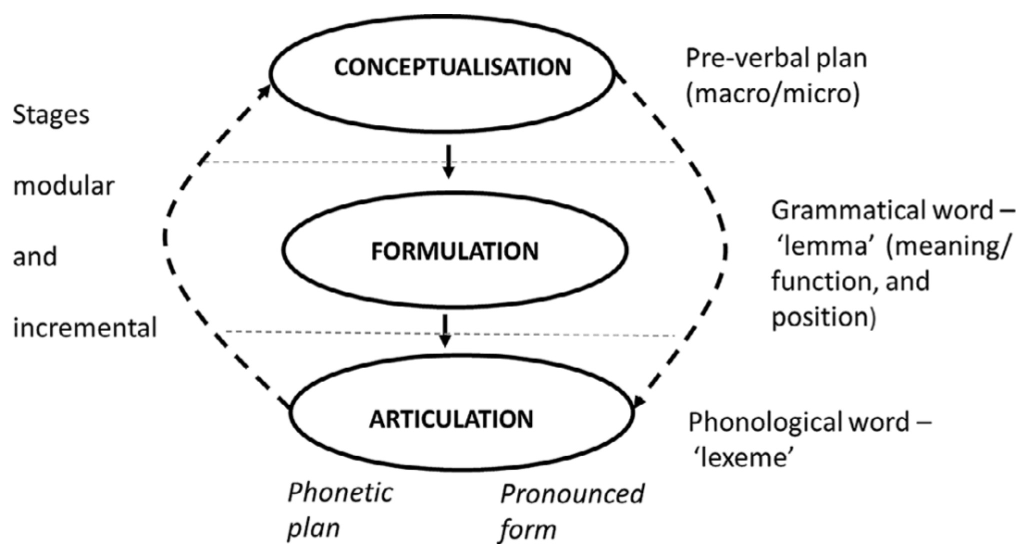

Figure 1. Illustration of Levelt's Speech Production Model (Tavakoli & Wright, 2020)

In summary, Levelt's (1989) model is widely accepted among cognitive SLA researchers as the most comprehensive model of L1 speech production, while it also accounts for many of the disfluencies that L2 speakers experience. Such disfluencies are thought to mainly occur at the formulation stage, especially when lexical-grammatical features can only be recalled via effortful and conscious processing, rather than automatic parallel processing (Kormos, 2006). What this model lacks, however, is an explanation of how speaking skills develop over time, as well as any indication of how the environment or interlocutors influence speech production and the cognitive processes that underpin it (Segalowitz, 2010).

Automaticity

Based on Levelt's (1989) model, the key to speaking fluently is parallel processing because this skill allows the different stages of speech production to operate simultaneously. For example, the ability to articulate one message while at the same time conceptualizing a subsequent one minimizes pauses and allows speech to proceed smoothly. Given the inability of humans to pay conscious attention to several things at once (Kahneman, 1973), parallel processing relies on the various stages of speech production becoming automatic, thereby requiring little cognitive effort.

Stating that a process has become automatic implies not only an increase in speed but also greater processing efficiency. In other words, underlying cognitive mechanisms are assumed to have been restructured such that there is less reliance on controlled supervisory and decision-making functions (Segalowitz & Segalowitz 1993). According to Segalowitz (2003), the characteristics of automaticity, in addition to speed, are that it is: ballistic, so it is impossible to stop the information being processed even with conscious effort; load independent, meaning that processing is unrelated to the amount of information involved; effortless, so that processing does not require attention or effort; and unconscious, meaning that processing is not encoded in conscious memory.

For L2 learners, fluency depends not only on the amount of lexical and grammatical knowledge they have, but crucially on automatizing access to that knowledge (Kormos, 2006; Suzuki & Kormos, 2023). Rule-based theories, such as

Anderson's (1992) Adaptive Control of Thought (ACT) theory, posit that automaticity results from the proceduralization, and ultimately automatization, of declarative knowledge. By this account, the first stage of skill acquisition involves obtaining declarative knowledge about how to perform the skill in question, such as playing a musical instrument, driving a vehicle, or learning a second language. The next stage—developing procedural knowledge—involves performing the skill by drawing on the consciously held declarative knowledge. During this stage, micro-elements of the skill are chunked into larger representations that can be accessed as a whole, thus reducing the burden on working memory. The final stage is automatization, whereby the skill can be performed without conscious effort and at which point the original declarative knowledge might no longer even be available for recall. In terms of language learning, this process explains why advanced speakers can use a grammatical form accurately even if they cannot remember the initial rule (Kormos, 2006). The key to this process is large amounts of deliberate practice, through which "knowledge that" becomes "knowledge how" (DeKeyser, 2007).

Although rule-based theories offer a convincing explanation of how consciously-held knowledge becomes automatized, not all learning involves the transition from declarative to procedural knowledge. According to Logan's (1988) instance theory, although novice learners initially use algorithms to complete tasks, single-step retrieval from memory eventually overwhelms slower algorithmic processing. During fluent speech, for example, the application of grammatical rules is outpaced by direct retrieval of the completed form, given sufficient practice and exposure to that form. As stated by Logan (1988), "Automatization reflects a transition from algorithm-based performance to memory-based performance" (p. 493). One problem with this theory, however, is that it does not account for how morpho-syntactic rules are generalized to novel situations for which learners have no memory representations to access (Kormos, 2006). It appears to be more clearly applicable to lexical knowledge, including formulaic language, which requires no rule to be processed and is retrieved on the basis of association with co-occurring items (Ellis, 2001). Therefore, although rule-based and instance-based theories both relate to automaticity neither type is comprehensive, as they arguably explain different aspects of the language-learning process. Similarly, Skehan (1998) argued that language processing relies on both rule- and memory-based retrieval, which he referred to as dual-mode processing. Nevertheless, all these theories offer support to the intuitive notion that L2 fluency development requires extensive practice of previously taught items (Nation, 2007).

Formulaic Language

The use of formulaic language has long been recognized as an important element of fluent speech production (Pawley & Syder, 1983). Although precise definitions vary, researchers agree that formulaic language consists of sequences of words that often occur together and are retrieved as a single item (Kormos, 2006). Wray and Perkins (2000) identified as many as 40 terms used in the literature to describe such sequences, including collocations, idioms, phrasal verbs, and fixed expressions. If such prefabricated chunks can be swiftly recalled from memory, as hypothesized in Logan's instance theory, cognitive burden is reduced and the parallel processing necessary for fluent production is facilitated. For example, the expression "I'm sorry to hear that" consists of several discrete syntactic and lexical elements and would therefore be difficult for a novice speaker to assemble piece by piece using rule-based processing. On the other hand, if such expressions are stored and retrieved as single items processing effort is reduced, allowing speakers to focus on other aspects of the communicative context (Wray, 2002).

This processing advantage was confirmed by Yan (2020), who used elicited imitation tasks to investigate the effect of formulaic sequences on speech rate and pausing among L1 and L2 speakers of English. Participants were asked to repeat sentences, half of which contained three- and four-word lexical bundles, such as "what I'm saying" and "have a question." Results revealed that, although overall speech rate was unaffected, formulaic sequences significantly reduced the number of silent pauses, especially in longer sentences and for L2 intermediate speakers. This finding, and the fact that both L1 and L2 speakers rarely pause in the middle of formulaic sequences (Tavakoli, 2011), suggests that they facilitate oral fluency.

There is currently a lack of research investigating how L2 learners acquire formulaic language, although in an exploratory study, Wood (2009) found that an intermediate Japanese learner of English was able to borrow formulaic sequences from models provided during lessons. The result was increased speech rate and mean length of run, as measured on a monologic narrative task performed after the treatment. The borrowed chunks also became more complex, including phrases such as "When I was a little girl..." and "The interesting thing is that..." In a subsequent study of 11 ESL students, Wood (2010) also found that the amount of formulaic language increased as learners became able to complete a narrative retelling task more fluently.

A positive relationship between formulaic language use and fluency was also demonstrated in a larger study of 56 L2 English learners during a high-stakes oral exam

(Tavakoli & Uchihara, 2019). The authors found a moderate correlation between articulation rate and the use of high-frequency multiword sequences (MWSs), which supports the idea that such sequences are stored and retrieved as individual items (Kormos, 2006). Furthermore, speakers who used a larger proportion of MWSs paused less mid-clause, which is associated with disfluencies related to lexical retrieval (Kahng, 2014). In pedagogical terms, if learners internalize such sequences to the point of automatization, attentional resources can then be diverted to other aspects of performance, such as monitoring and articulation, resulting in more fluent output (Kormos, 2006).

Researching L2 Speaking Fluency

Repeated Monologue Activities

Fluency development has traditionally been researched on an individual basis, often using pedagogic activities in which speakers are asked to repeat a monologue under increasing time pressure. One such technique is the 4/3/2 activity, in which speakers attempt to repeat the same content within a reduced time frame—four minutes, then three minutes, then two minutes. This activity creates positive conditions for promoting fluency because, with the need for content generation removed after the first iteration, the speaker can focus entirely on delivering the content at a faster rate.

Nation (1989) used the 4/3/2 activity in a landmark exploratory study, in which six participants recalled a personal experience. Only two fluency measures were used: number of words per minute; and the combined number of hesitations, repetitions, and false starts per 100 words. However, using words per minute, rather than syllables, only gives an approximate measure as the number of syllables that words contain is variable.

Furthermore, the threshold of pause duration was not stated. Despite these limitations, the results showed that speech rate increased significantly between the first and third iterations, whereas the number of hesitations, false starts, and repetitions significantly decreased. Although this study played an important role in popularizing the 4/3/2 activity in the literature, methodologically it has been much improved upon since.

A variant of the 4/3/2 activity, using a more fine-grained methodology, was adopted by Thai and Boers (2016). Speech rate was calculated as the mean number of syllables per minute, with both raw and trimmed measures presented—the latter referring to syllables per minute after filled pauses, repeated syllables, and false starts have been excluded. A third measure of fluency was nonphonation/time ratio, which refers to the proportion of silent pauses per minute, with the pause threshold set at 300ms. However, the most important innovation was the inclusion of a comparison group, which allowed the effects of the reduced time condition to be contrasted with straightforward task repetition.

Results showed that the reduced-time (3/2/1) condition produced a much higher rate of speech than the constant-time (2/2/2) condition, although syntactic accuracy and complexity improved in the latter condition only. These findings are consistent with the idea that increasing time pressure promotes fluency development, although the authors claimed that these findings support Skehan's (2009) Trade-off Hypothesis, which states that increases in fluency come at the cost of accuracy and/or complexity. However, this conclusion is debatable, as there were no significant changes in either complexity or accuracy in the reduced-time condition. Overall, despite the relatively small sample size of 20 students, and the adapted nature of the activity (i.e., 3/2/1, rather than 4/3/2), this study has contributed to the growing body of evidence that reduced-time repeated monologues are an effective way to promote spoken fluency, especially rate of speech, at least in the short-term.

In psycholinguistic terms, the positive effects of task repetition on fluency are related to proceduralization and eventual automatization (DeKeyser, 2007). Repeating lexical and syntactic items has the effect of gradually embedding them in the learner's mind in association with the function they are used to perform. Such items then become easier to retrieve when called for in similar situations in the future. This process helps to explain the importance of formulaic speech for fluency development, as prefabricated phrases that are available for rapid use allow speakers to assemble units of discourse more quickly than if each phrase has to be reassembled item by item (Wray, 2002).

Using this theoretical justification, De Jong and Perfetti (2011) compared a group that talked about the same topic for each round of 4/3/2 with a group that was given a different topic each time. Results showed that students in the repetition group made improvements in long-term fluency, as measured on delayed posttests, which crucially transferred to new topics. Specifically, the repetition group made gains in terms of length of fluent run (i.e., number of syllables produced between pauses) and reduced length of pauses, whereas those in the changing topic group did not. The authors interpreted these results to mean that the repeated use of linguistic forms had precipitated changes in underlying cognitive processing mechanisms (i.e., proceduralization), resulting in long-term improvements in fluency. Moreover, few topic-related words from the 4/3/2 training activities were repeated on the posttests, indicating that, rather than the retrieval of specific lexical items, long-term proceduralization concerned sentence structure more generally. It also suggests that improvements made under these circumstances might be transferrable to novel topics.

Measuring Fluency During Interaction

As mentioned above, most studies measuring L2 fluency have primarily been focused on monologic speech performance. There are clear methodological advantages to measuring fluency in this way, as extended stretches of speech can be produced without overt interference from an interlocutor thereby facilitating a more predictable form of output (Tavakoli & Wright, 2020). Moreover, this approach accords with Levelt's (1989) three-stage model, which is based on the cognitive processes underlying individual speech production. In contrast, interaction-based tasks produce less predictable and more fragmented output. They also produce interaction-specific features, such as inter-turn gaps or overlaps, which are unattributable to any individual (Tavakoli, 2016).

Nonetheless, some researchers have attempted to investigate L2 fluency using dialogue tasks. In a study of 64 learners of L2 Dutch, Michel (2011) found that a telephone conversation task elicited more fluent output than a monologic task that required leaving a message on an answering machine. Specifically, the dialogue activity produced a faster rate of speech, as well as fewer pauses and repairs, than the monologue activity for the L2 learners. The difference between the two modes reached statistical significance for all measures, with repair and pausing producing larger effect sizes (partial $\eta^2 > .25$) than speech rate. The author explained these results by arguing that dialogues are cognitively less demanding than monologues because interaction allows additional planning time during an interlocutor's turn, while also providing opportunities for sharing linguistic items, thus freeing up attentional resources (Michel, 2011).

Similarly, in a study of 35 EAP students, Tavakoli (2016) found that dialogues generated more fluent output than monologues in terms of speed, repair, and length of pauses, although not number and location of pauses. The author suggested that the collaborative nature of interaction promotes fluency not only by providing more planning time, but also by encouraging participants to consider the needs of their interlocutors. This accommodation therefore results in faster speech, as well as fewer repetitions and hesitations, in an attempt to facilitate smoother interaction (Tavakoli, 2016).

Although traditionally viewed as an individual attribute, fluency during interaction is co-constructed, as participants alternate between speaker and listener roles (Kasper, 2006). Moreover, McCarthy (2010) coined the term *confluence* to highlight the fact that successful communication relies on speakers relating their comments to previous turns, rather than producing a chain of disconnected comments. However, in terms of methodology, Tavakoli (2016) pointed out that the interaction-specific features of fluency have yet to be systematically investigated. She also questioned whether fluency represents the same construct in monologue and dialogue performance, given that the speaker's attention is additionally focused on the needs of the interlocutor during interaction. Moreover, Peltonen (2017) argued that some traditional indicators of disfluency, such as repetitions, filled pauses, and lexical fillers, are positively utilized during interaction to avoid silence and hold the floor. Viewing these micro-elements as "potentially fluency-enhancing resources" (p. 2) used to maintain the interactional flow underlines the need to examine fluency in social, as well as individual, settings. There has been a lack of such research to date, perhaps due to the methodological complications mentioned previously; however, in this study I measure aspects of fluency in both monologue and interactive activities.

Perceived Fluency

Interaction, of course, requires interlocutors as well as speakers. Perceived fluency, defined by Segalowitz (2010) as the "inferences listeners make about a speaker's cognitive fluency based on their perception of utterance fluency" (p. 48), constitutes the third element of his widely used triadic framework. The way in which speech is perceived by listeners is an essential aspect of fluency in a holistic sense, as it brings together all three elements of this framework.

Several researchers have attempted to relate objective measures of utterance fluency to subjective ratings of perceived fluency, thus reinforcing the intertwined nature of these elements. A groundbreaking early attempt was made by Riggenbach (1991), who found that speakers rated as highly fluent spoke faster and with fewer pauses than those rated as lacking fluency. Subsequent studies have confirmed the relationship between utterance and perceived fluency—regardless of whether the raters were teachers, experts, untrained, native or nonnative speakers—with pausing phenomena and speed of delivery consistently found to predict raters' perceptions (De Jong et al., 2013; Suzuki et al., 2021). For example, Suzuki and Kormos (2020) found that frequency of unfilled mid-clause pauses was most strongly associated (i.e., negatively correlated) with perceived fluency. This finding is important because such pausing reflects fluency development more generally, as L2 learners often pause longer and more frequently than L1 speakers mid-clause, but not at clause boundaries (Tavakoli, 2011). With reference to Levelt's (1989) model, Skehan et al. (2016) hypothesized that clause-end pausing indicates pre-linguistic planning at the conceptualizer stage, whereas mid-clause pausing is associated with linguistic micro-planning, and hence disfluency at the formulator or articulator stages.

In the above studies, perceived fluency has been related to traditional measures of utterance fluency, that is, speech rate, breakdown, and repair. However, in an innovative study, Van Os et al. (2020) related perceived fluency to turn-taking behavior during interaction. Given that speakers of various languages tend to avoid overlaps and minimize gaps between turns (Stivers et al., 2009), the authors had raters judge samples of Dutch that had been manipulated to include these features. They found that overlapping caused native speakers to be negatively rated, whereas extended gaps between turns caused non-native speakers to be negatively rated. In other words, fast speech with overlaps was seen as too fast, whereas slow speech with gaps was seen as too slow. Due to the effect of turn-taking behavior on fluency ratings, the authors argued that Peltonen's (2017) term *dialogue fluency* should be adopted as an additional category of utterance fluency when measured during interaction. They also argued that turn-taking, as an aspect of fluency, should be reflected in language pedagogy and assessment. By highlighting the co-constructed nature of fluency in interaction, not only has this study broadened understanding of the construct, it has also bridged the gap between fluency and interactional competence research.

Subjective ratings of fluency have traditionally been measured using numerical rating scales (Suzuki & Kormos, 2020). However, the scores obtained are inevitably influenced by the instructions given. For example, if raters are instructed to pay attention to pausing and speech rate, these elements are more salient and therefore more likely to be reflected in the results. On the other hand, if no instructions are given, raters simply define the construct in their own—perhaps idiosyncratic—way, thus undermining the validity of the results (De Jong et al., 2013). In other words, what raters perceive as fluent (or disfluent) is not necessarily related to the speaker's underlying cognitive ability. To counter this threat, some researchers have provided an analytic rating scale, with descriptors that clarify what the raters should focus on when making their judgments (e.g., Nitta & Nakatsuhara, 2014). This approach is adopted in the current study.

Taking an even wider perspective, it has been argued that fluency forms part of overall communicative adequacy, defined as "the degree to which a learner's performance is more or less successful in achieving the task's goals efficiently" (Pallotti, 2009, p. 596). For example, speech that is superficially fluent but irrelevant or difficult to comprehend, whether grammatically accurate or not, cannot be described as functionally adequate. Despite this fact, few researchers have reported whether a task's communicative aims were achieved. In one study that did address this issue, Révész et al. (2016) assessed the impact of fluency, as well as accuracy and complexity, on ratings of communicative adequacy in a variety of communicative situations. Raters assessed whether speakers were easy to understand, delivered the message effectively and appropriately, and provided sufficient detail. Fluency was found to be the strongest predictor of communicative adequacy—regardless of task type—with frequency of filled pauses alone explaining 15% of the variance; that is, speaking was rated higher if it contained fewer filled pauses. No other significant predictor accounted for more than 7% as an individual factor. These results support the authors' assertion that the dominance of lexical and grammatical issues within SLA is unwarranted, especially in communicative settings, as messages can be conveyed without the use of complex or completely accurate language. In the current study, raters scored participants on their communicative adequacy, as well as fluency, during the interactive activities.

The Effect of the Learning Context

Another approach to researching fluency development is to investigate the role of context, with researchers often comparing development across different contexts. For example, Segalowitz and Freed (2004) investigated differences in cognitive and utterance fluency among 40 students studying Spanish, either through study abroad or a foreign-language course at their home university. Utterance fluency data were obtained from oral proficiency interviews, while cognitive fluency was measured on reaction time tests of word recognition. Results at the end of one 13-week semester showed that although both groups made improvements in cognitive fluency, only the study abroad group had made significant gains in oral proficiency and utterance fluency. The study abroad students were enrolled in more language classes and had far greater opportunities to use the language than the domestic students, but the authors stated that cognitive readiness, as shown on word recognition pretest scores, rather than the amount of language contact was the decisive factor, given the short-term nature of the program. In other words, only those students who had reached a threshold of cognitive fluency prior to departure were able to take advantage of the communicative opportunities on offer during study abroad.

In another single-semester study, this time with learners of French, the same researchers contrasted fluency development in three contexts: domestic foreign language study, study abroad, and a domestic immersion program (Freed et al., 2004). Contrary to the authors' expectations, the immersion students made greater utterance fluency gains than the study abroad group, despite the fact that they had a similar number of classroom hours. The foreign language study group made no such gains at all, even though communicative learning was emphasized in the curriculum. The authors explained this result in terms of time spent using the target language outside of class, which had also been recorded. The immersion students used the language significantly more than those studying abroad, who tended to use their native language with fellow students, despite the greater opportunities for target language use afforded by the environment. In conclusion, the authors stressed:

> it is not the context per se that promotes various types of learning but rather… the nature of the interactions, the quality of the experiences, and the efforts made to use the L2 that render one context superior to another with respect to language gain (Freed et al., 2004, p. 298).

Rather than the primacy of context, therefore, these findings support arguments in favor of skill acquisition theory (Anderson, 1992; DeKeyser, 2007) and time on task in terms of fluency development.

Another comparative study, using a variety of CAF measures to contrast the effect of study abroad and EFL contexts, was conducted by Mora and Valls-Ferrer (2012). The researchers investigated a group of 30 students over a two-year period, comparing any initial development made in the EFL context with subsequent gains made after a three-month period of study abroad. No improvements were observed during the EFL period, whereas robust fluency gains, as well as moderate gains in accuracy, were observed following the study abroad program. The authors concluded that improvements in mean length of run and pause duration indicate that during the study abroad period the participants had proceduralized previously learned explicit linguistic knowledge.

In order to understand these differences in outcome, however, it is necessary to examine the nature of instruction and opportunities for communication provided in both contexts. Importantly, the authors stated that during the EFL period there was no specific focus on oral communication skills and that instruction largely consisted of traditional grammar instruction. Furthermore, the learners had limited, if any, opportunities for L2 use outside the classroom. Given these circumstances, it is hardly surprising that no fluency gains were observed. In contrast, during the period of study abroad, there were many opportunities for communicative L2 interaction beyond the classroom. The results of this study therefore emphasize the influence of context by highlighting the learning opportunities provided in each type, although previously mentioned arguments about time on task and cognitive readiness (Segalowitz & Freed, 2004) must also be borne in mind.

Indeed, a lack of opportunities for spoken production has been a common theme in foreign language teaching contexts. A limited amount of instruction time, combined with limited L2 input, peer interaction, and use outside the classroom (Muñoz, 2008), means that such settings are often unsuitable for the development of oral proficiency. Nevertheless, fluency gains have been reported in EFL contexts, even within a limited time frame. In one such study, Hanzawa (2021) recorded 50 Japanese university students delivering a personal narrative at three points during an academic year. Although no development was observed regarding articulation rate or pause frequency, length of pauses declined significantly. The author suggested that this result reflected the early stages of fluency development and that it would take longer for pause frequency to additionally decline. Moreover, although between-clause pause length had only declined by the end of the academic year, within-clause pause length fell during both semesters. In addition, Doe (2025) reported significant gains in phonation/time ratio and mean length of pause during 3/2/1 activities recorded over a single semester, also in a Japanese university. Furthermore, although the gains were moderate, they were detected by raters using a five-point analytic scale. In sum, these studies suggest that aspects of fluency can improve during one semester of study, even in an EFL environment, as is the case in the current study.

Pedagogical Interventions

Given the reported disparity between fluency development in EFL and study abroad contexts, the question arises as to whether it can be bridged by pedagogical interventions. Based on theories of fluency development, several pedagogical frameworks have been suggested by researchers (e.g., Gatbonton & Segalowitz, 1998; Nation, 2007). However, the practical implications of fluency research are rarely reflected in teaching resources, syllabi, or the activities found in commercially available textbooks (Rossiter et al., 2010; Tavakoli, 2023).

In his influential four strands pedagogical approach, Nation (2007) suggested that roughly one quarter of instruction time should be devoted to activities specifically aimed at fluency development. These activities should be meaning-focused and should incorporate an element of time pressure so that learners are encouraged to perform the skill faster than they were previously able to. Activities should not include any unknown lexical or grammatical items as the intention is to proceduralize existing knowledge, rather than learn new items, at this stage. The previously mentioned 4/3/2 activity is an archetypal example of a speaking activity aimed at fluency development.

The other three strands of a balanced language curriculum consist of meaning-focused input, meaning-focused output, and language-focused learning, of which only the last prioritizes a deliberate focus on linguistic form. Classroom conversations and discussions come under the domain of meaning-focused output, as the emphasis is on conveying meaning through largely familiar language, but without the added time pressure element that fluency development activities require. Nevertheless, such activities should incidentally improve fluency—if sufficient opportunities for output are provided—as learners become more comfortable productively using linguistic items they have previously studied. According to Nation, each strand should account for roughly 25% of the curriculum. Although he conceded that this number is somewhat arbitrary, the emphasis on fluency development and productive language use holds great appeal for language teachers working within EFL contexts such as Japan, in which students experience heavily form-focused instruction during formal education prior to university (Koizumi et al., 2022; Nishino & Watanabe, 2008).

An earlier attempt to create a theory-based framework for fluency development activities can be found in Gatbonton and Segalowitz's (1988) work on creative automaticity. The authors specified five principles to be considered when designing activities that aim to promote automaticity, which are that activities should: be genuinely communicative; incorporate aspects of authentic communication, such as time pressure; focus on real-world language functions, such as describing or giving directions; require the use of short formulaic utterances; and be inherently repetitive, such that the activity can be repeated with different partners. The authors emphasized the use of prefabricated chunks of language in these activities, given that "memorized clauses and clause-sequences form a high proportion of the fluent stretches of speech heard in everyday conversation" (Pawley & Snyder, 1983, p. 208).

This framework was built on and developed in the same authors' ACCESS fluency development methodology (Gatbonton & Segalowitz's, 2005), in which a free communication phase was added to the more controlled practice activities in the original model. This free communication phase relates to the concept of transfer appropriate processing (TAP) because if information is retrieved from long-term memory more easily in situations resembling those in which it was acquired, classroom activities should reflect real-world communication as much as possible (Lightbown, 2008). In other words, using language in unstructured situations outside of class will likely prove difficult if practice in class consists solely of highly structured and predictable activities.

An innovative approach to fluency development was investigated by Tavakoli, et al. (2016), who compared the effects of awareness-raising activities and fluency strategy training with a control group focused on the general development of listening and speaking skills. This training included metalinguistic and awareness-raising activities, such as recording and evaluating the fluency of learners' speech, and practicing lexical fillers (e.g., "Let me see..."). The treatment group made significant gains on four of the nine fluency measures over a four-week period with rate of speech having the largest effect size (partial $\eta^2 = .40$). The authors suggested that a combination of declarative knowledge, facilitated by the awareness-raising activities, and proceduralization, afforded by the practice activities, was probably responsible for these gains, and therefore that pedagogical interventions can influence fluency development. However, as the authors pointed out, it would be useful to ascertain whether these gains were sustained beyond the short-term duration of the study.

Summary of L2 Speaking Fluency

The traditional psycholinguistic view, reflected in Levelt's (1989) model and the use of monologue data, is that fluency is a component of individual L2 proficiency. This idea is related to theories of skill acquisition and the need for extensive practice to proceduralize, and gradually automatize, linguistic knowledge (DeKeyser, 2007). However, the realization that fluency is a multi-faceted construct, and exists at least partly in the ear of the beholder (Freed, 2000), has led to broader definitions that include the element of listener perception in addition to learner-internal elements (Segalowitz, 2010). This development has led some researchers to attempt to relate objective measures of fluency to subjective measures, whereas others have measured aspects of fluency during interaction. Moreover, the notion that fluent interaction is a co-constructed achievement that has to be collaboratively maintained highlights the fact that fluency has a sociolinguistic, as well as a psycholinguistic, dimension. Pedagogical frameworks have been devised that account for the role of context and the requirements of transfer appropriate processing, while promoting automatization in communicative settings (Gatbonton & Segalowitz, 2005). These principles form the foundation of the pedagogical intervention underpinning the current study, in which fluency is measured in individual and interactive contexts, both objectively and subjectively.

Interactional Competence

Originally coined by Kramsch (1986), the term *interactional competence* refers to the ability to manage turn-taking and problems of understanding (Wong & Waring, 2021), and has become widely used to emphasize the importance of the social context in which language is typically used. Interaction is inherently co-constructed in that each contribution must be understood in terms of how it relates to the ongoing discourse and how it builds upon previous contributions, while also providing the context for subsequent contributions (Salaberry & Burch, 2021). According to Young (2008), interactional competence refers to the relationship between participants' resources and the context within which they are utilized. These resources are not only linguistic (e.g., grammar and vocabulary) and interactional (e.g., turn-taking and repair), but also the context-specific identities (e.g., novice or expert) that participants adopt during the interaction.

The idea that language use is inherently social recalls Canale and Swain's (1980) application of communicative competence to SLA, and its focus on the knowledge—grammatical, sociolinguistic, discourse, and strategic—required for successful interaction. However, constructivist researchers (e.g., He & Young, 1998) have criticized its emphasis on the static acquisition of knowledge within the individual, rather than the dynamic co-construction of knowledge during interaction According to Young (2011):

> the fundamental difference between communicative competence and IC is that an individual's knowledge and employment of these resources is contingent on what other participants do; that is, IC is distributed across participants and varies in different interactional practices. And the most fundamental difference between communicative and interactional competence is that IC is not what a person *knows*, it is what a person *does* together with others (p. 430).

In other words, an emphasis on context and the joint performance of all participants is what distinguishes interactional competence from communicative competence.

In practical terms, however, the key question is the extent to which the resources that participants bring to one context can be redeployed in others. If, as argued by He and Young (1998), interactional competence is jointly constructed by participants making use of their shared resources, it cannot therefore be attributed to any individual (Young, 2011). This perspective presents an obvious dilemma in terms of generalizability, as it suggests that the construct is entirely local and context-specific. Moreover, Bachman (2007) has pointed out the contradiction between conceptualizing interactional competence as co-constructed and distributed among participants, but also consisting of resources brought by individuals. In an attempt to reconcile this apparent contradiction, Lam et al. (2023) advocated "a hybrid individual/social view, which recognises interaction as a co-constructed, shared achievement that is enabled by individuals' linguistic and embodied resources." (p. 14).

Furthermore, the notion that interactional competence is disembodied from the individuals participating in the interaction has been rejected by Fulcher (2010), who argued that this view conflates an individual's ability to communicate with the act of communication itself. According to Fulcher (2010):

> No one would wish to deny that each instance of communication, each instance of new discourse, arises within a context and is dependent upon the variables at play. But this does not mean that the competence to engage successfully in these new interactions is also newly generated in each instance of communication (p. 110).

I believe that this account offers more hope to practitioners and learners alike that the ability to interact, so fundamental to L2 speaking proficiency, can be transferred to novel contexts consisting of different participants. Moreover, viewing interactional competence primarily as a suite of abilities that each individual brings to the interaction, rather than something entirely co-constructed and ephemeral, makes the construct measurable and allows its development to be tracked over time for pedagogical or assessment purposes. This perspective also informs the current study, as the aim is to measure and track interactional competence longitudinally.

Nevertheless, focusing on the resources of the individual does not mean neglecting the role of context. On the contrary, familiarity with aspects of a particular context—such as activity type, participant role, or conventional behavior (Hall & Pekarerk Doehler, 2011)—is bound to aid individual performance, reflecting the complex relationship between variable context and more stable ability attributes (Chalhoub-Deville, 2003). In other words, the more familiar learners are with a given context the more efficiently they can activate their latent knowledge. However, the contingent nature of interactional competence means that it is only observable during interaction (Roever & Kasper, 2018); therefore, how well an individual can demonstrate their competence depends to a large extent on how a specific interaction unfolds and how the other participants behave. Lam (2018) used the analogy of a champion tennis player being unable to demonstrate the full range of their skills when playing against a novice who is unable to return the ball, as opposed to a fellow professional. In terms of language testing, studies by Roever and Kasper (2018) and Ross (2018) have shown how oral examiners shape interaction in ways that constrain candidates' opportunities to demonstrate their interactional competence.

The fact that interactional competence is contingent on the behavior of others raises the issue of whether it can be measured reliably. One way to address this point in an individual testing situation is to tightly control the contributions of the interlocutor (i.e., the examiner). Adopting this principle, Van Batenburg et al. (2018) found that by using tasks designed to elicit specific communicative strategies, such as including unfamiliar words in the examiner's script, they were able to achieve high levels of agreement among raters regarding candidates' performance. The authors concluded that these high correlations—observed across all six tasks—suggest that interactional ability can be measured reliably and can therefore be regarded as an individual trait. Nevertheless, in the case of freer speaking activities there is no guarantee that specific aspects of interactional competence will be manifested. For example, Kley at al. (2021) found that, during assessment tasks designed to elicit spontaneous talk with an L1 speaker, American university learners of L2 Chinese did not use some of the repair practices they had been taught. The authors speculated about possible reasons, such as mutual understanding not being threatened or potential misunderstandings being preempted by the interlocutor. To sum up, although researchers disagree over the extent to which interactional competence pertains to the individual, what can be observed about one's interactional ability is undeniably co-constructed (Lam, 2018). This fact makes measuring and assessing the construct a challenge in all but the most tightly controlled contexts.

The competing views of interactional competence outlined above make operationalizing the construct a challenge, especially because, like fluency, it can be conceptualized in both broad and narrow terms. According to Galaczi and Taylor (2018), on a macro-level interactional competence can be defined as "the ability to co-construct interaction in a purposeful and meaningful way, taking into account sociocultural and pragmatic dimensions of the speech situation and event." (p. 226). Pallotti (2019), meanwhile, has defined interactional competence as "that competence which is specifically needed to inter-act, that is, to participate in courses of action involving two or more speakers, and which crucially has to do with turn-taking dynamics" (pp. 6–7). This definition is more applicable to the current study as it is based on one specific, largely monocultural EFL context; therefore, pragmatic and sociocultural factors are likely to be greatly reduced. In other words, the fact that the students are familiar with this particular classroom environment means that the focus of development is on turn-taking elements, rather than issues of politeness and formality, for example.

In its narrower sense, interactional competence can be divided into a range of skills used to accomplish various social actions, such as turn-taking and repairing breakdowns in communication. Galaczi and Taylor (2018) listed these skills as topic management; turn management; interactive listening; breakdown repair; and non-verbal or visual behavior, which they presented in the tree illustration shown in Figure 2 to represent the organic nature of spoken interaction.

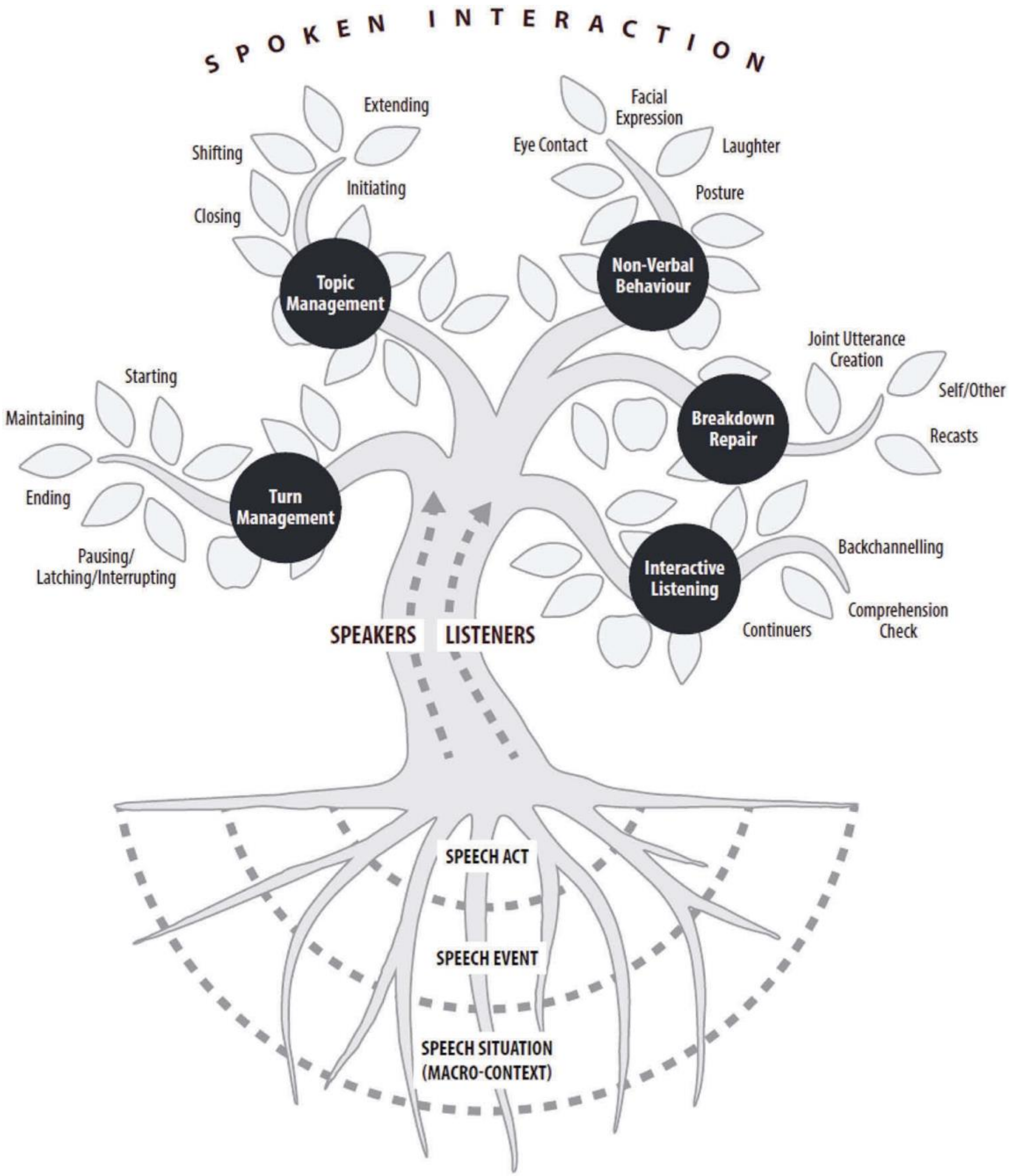

Figure 2. <u>Defining Interactional Competence (Galaczi & Taylor, 2018)</u>

The roots of the tree represent the macro-level context of the speech situation, as well as the micro-level context of the speech event and the individual speech act. The trunk represents the participants in the interaction, whether dyadic or group-based, whereas the skills listed above are represented by the larger branches. Finally, the smaller branches represent micro-features of each skill; for example, interactive listening consists of what the authors call backchannelling, comprehension checks, and continuers. Not only does breaking interactional competence down in this way provide a more practical and comprehensive definition, it is also a necessary precursor to assessing or measuring the construct. All the features mentioned above—topic management, turn management, interactive listening, breakdown repair, and non-verbal or visual behavior—are relevant to the current study because they are all fundamental aspects of group-based interaction (see Chapter 3 for an explanation of how these skills have been adapted for coding and analysis).

L1 versus L2 Interactional Competence

A key concept in interactional competence research is that of intersubjectivity. To communicate successfully, participants must not only share sufficient knowledge of the world and the external context, but they must also construct a "shared internal context" (Kramsch, 1986 p. 367) by negotiating, anticipating, and clarifying intended meaning. Intersubjectivity, therefore, is "the conscious attribution of intentional acts to others and involves putting oneself in the shoes of an interlocutor" (Young, 2011, p. 430). In other words, communication, at its most fundamental, represents the attempt to transfer the contents of one mind to another, which presupposes an innate ability to understand what another person is thinking. This concept clearly precedes L2 learning and is the foundation of what Levinson (2006) has called the "interaction engine," which enables learners to draw on universal pre-linguistic communication abilities when interacting in an L2. For example, Carroll (2004) has shown that L2 learners often establish recipiency, that is, the readiness of an interlocutor to receive talk, by waiting for gaze alignment before beginning a speaking turn. This kind of feature is independent of language or proficiency level, yet it forms part of the turn-taking system that studies have shown is remarkably similar across unrelated cultures (Levinson, 2022).

However, acknowledging that the ability to interact precedes L2 learning presents a dilemma because, as several researchers (e.g., Kasper, 2009) have pointed out, it implies that interactional competence is both the object and resource of learning. In other words, social interaction not only shapes the learning process but is shaped by it (Pekarek Doehler, 2018). In order to resolve this ambiguity, different labels have been proposed to distinguish between innate and learned forms of interactional competence. For example, Kecskes et al. (2018) suggested the term *basic interactional competence* (BIC) to describe features that begin developing in pre-linguistic infancy and *applied interactional competence* (AIC) for culture-specific features. Alternatively, Hall (2018) referred to the underlying form as universal interactional competence and interactional repertoires for the objects of L2 learning, which include the micro-skills proposed by Galaczi and Taylor (2018). Learning new skills and techniques, thereby diversifying these repertoires, enables learners to adapt to local circumstances and deal with a wider range of interactional contexts, which is arguably the key to the development of L2 interactional competence (Pekarek Doehler & Pochen-Berger, 2015). These repertoires or micro-skills—the objects of L2 learning—are what is referred to as interactional competence in the current study. They include features such as follow-up questions, comprehension checks, and backchannels (see Chapter 3 for a full inventory of the interactional features adopted in this study).

Researching Interactional Competence

Interactional competence researchers have typically focused on one micro-skill, using conversation analysis (CA) techniques to highlight its use either in a naturalistic or classroom environment. Using CA to explain L2 learning shifts the focus away from abstract linguistic structures such as verb tenses or morphemes—which have traditionally been prioritized in SLA—and towards the processes of interaction, such as turn-taking and sequence organization (Hellermann, 2009a). As the current study is classroom-based, the studies reviewed in this section are limited to that context.

In one such study, Hellermann (2011) examined instances of two novice learners of English either directly repairing their partner's talk, or initiating such repair, over a 50-week period. By later time points, the participants had expanded their repertoire of methods for doing other-initiated repair to include such techniques as spelling out a trouble-source word and providing alternative examples. Furthermore, the objects of repair expanded from mainly lexical items in the initial stages to broader aspects, such as orientation to a role-play task, by the end of the study. It is noticeable, however, that even in the beginning stages of learning, the participants attempted peer repair at the earliest available opportunity (i.e., in the immediately subsequent turn) indicating that the transfer of an underlying L1 competence was taking place.

In another such study—in this case cross-sectional—Pekarek Doehler and Pochon-Berger (2011) found that advanced secondary-school learners of French demonstrated a wider range of techniques for expressing disagreement than lower-intermediate learners. The advanced learners regularly used 'yes-but' constructions, in which the disagreement was delayed until later in the turn, while also mitigating them with accounts and explanations. This kind of elaboration mirrors L1 discourse, in which disagreement—as a dispreferred action—tends to be downgraded or prefaced by an initial agreement token (Pomerantz, 1984). The lower-intermediate learners, by contrast, were more likely to express strong disagreements using simple polarity markers such as ‘no’. This finding supports that of an earlier study by Bardovi-Harlig and Salsbury (2004), who observed that adult learners of various L1 backgrounds progressed from short, direct disagreements to more elaborated forms during an intensive ESL course. This development occurred without any direct instruction in pragmatics, although the authors suggested that it could be enhanced by explicit teaching. It is important to note, however, that in both studies rudimentary methods of disagreement were not simply replaced by more sophisticated forms at later stages of development. Rather, more advanced learners continued to produce direct, turn-initial disagreements at times, presumably when they judged it appropriate to do so, indicating an expanded repertoire of resources.

This finding is congruent with Hellermann’s (2009b) study showing that unmitigated ‘no’ tokens, which might be considered inappropriate in some contexts, are oriented to as appropriate by participants in a language-learning task. For example, they can be used for overt correction or to initiate peer repair, so that misunderstandings are not drawn out over several turns. Therefore, learners who are able to use various forms of disagreement—both direct and mitigated—have diversified their available resources for accomplishing this particular skill. It is this diversification that allows learners to adapt to the local contingencies of interaction in context-sensitive ways (Pekarek Doehler & Pochon-Berger, 2011).

Mining the same extensive data set as in his previously mentioned (2009b, 2011) studies, Hellermann (2008) also analyzed how learners at different proficiency levels approached task openings. He found that novice learners tended to launch directly into the activity, whereas intermediate learners were more likely to engage in prefatory talk such as confirming their co-participants' readiness to begin. This prefatory talk occurred even if there had been no indication of any trouble understanding the task, which Hellermann interpreted as participants "establishing interpersonal alignment for face-to-face interaction" (p. 64), rather than any perceived need to clarify instructions. In other words, the participants were establishing their shared status as learners within a community of practice. The use of prefatory talk by higher-proficiency learners when beginning a task recalls the tendency of such learners to mitigate their disagreements and push them further back in the speaking turn (Pekarek Doehler & Pochon-Berger, 2011). These kinds of elaborations, as opposed to the more abrupt forms of communication observed at lower levels, entail the ability to recipient-design talk and demonstrate sensitivity to the interactional context, which Pekarek Doehler (2019) has argued is central to L2 interactional competence development.

Another key index of interactional competence is the ability to take the initiative during interaction, with research showing that higher-proficiency learners tend to be more proactive, whereas lower-proficiency learners are more passive and often wait to be invited to participate (Pallotti, 2019). For example, Young and Miller (2004) investigated an adult ESL learner during a series of weekly writing conferences and observed the learner moving from peripheral to fuller participation. After only four weeks, the learner was able to perform many of the acts the instructor had initially performed, such as identifying problems with the writing and suggesting revisions. Despite claims that this kind of situated learning is practice-specific, Hellermann (2007) has argued that strategies used in one learning context, such as procuring an interlocutor, can be observed in similar contexts and therefore might be applicable to situations outside the classroom.

As well as managing one's own talk, being proactive also applies to how learners attend to the contributions of others. When investigating the development of Turkish undergraduate learners of L2 English, Sert (2019) observed extended inter-turn silences and few examples of active listening, such as tokens of acknowledgement, confirmation, or agreement, early in the course. However, by the end of the semester, the participants were displaying more attentive listening and collaborative turn sequences, such as elaborating on previous speakers' comments following confirmation tokens. In addition, the length of inter-turn silences was reduced and more lexical offers were made to assist the speaker. For example, on one occasion when a speaker appeared to be struggling to express an idea, another participant suggested the word *experience*, which was then taken up by the speaker in the following turn.

To sum up, researchers face several major challenges when investigating the development of interactional competence. These challenges include how to choose relevant units of analysis, as well as how to disambiguate observed changes in performance that are due to learner development from those that are due to changes in the local context of interaction (Hall & Pekarek Doehler, 2011). Nevertheless, taken together, the studies reviewed in this section suggest that L2 interactional competence development entails a diversification of techniques, with learners at higher levels of proficiency able to incorporate more elaborate forms of communication, as well as abrupt or direct forms, into their repertoires. This diversification, in addition to facilitating more proactive and fuller participation, provides learners with the flexibility to adapt to the local context of interaction while taking into account the needs of co-participants. However, given the classroom-based nature of these studies, a key question remains the extent to which such techniques can be explicitly taught and practiced in classroom environments (Pekarek Doehler, 2019).

Quantifying Interactional Competence

It is uncontroversial to state that some form of qualitative analysis is necessary to analyze the micro-features and patterns of interaction to observe how language is used "resourcefully, contingently, and contextually" (Firth & Wagner, 1997, p. 296). However, whether quantitative methods are also appropriate for researching interactional competence has proven far more controversial. Some researchers are skeptical, arguing that the highly contingent nature of interaction makes quantification of such features reductive. For example, Schegloff (1993) pointed out that whether an interactional feature, such as backchanneling, is produced or not depends not only on the ability or willingness of the speaker to produce it, but also on its relevance in that particular context. In fact, withholding a feature would demonstrate greater interactional competence than using it inappropriately or at an irrelevant juncture, although this kind of competence would not be revealed by statistical data. Other researchers, working with L2 data, have maintained that quantification can complement qualitative analyses to illustrate the development of interactional competence, assuming that more competent speakers exhibit different patterns of interaction from their less competent peers (Galaczi, 2008). In other words, using quantification to highlight trends, involving multiple participants, helps to address the thorny issue of how to disentangle individual contributions from a co-constructed format, given the contingent nature of interaction. Tracking the proportion of specific interactional moves learners make and how this proportion changes over time, therefore requires the kind of systematic analysis provided by the quantification of extensive data sets (Wagner et al., 2018).

Most attempts to measure interactional competence to date have come from the perspective of language assessment. For example, using data from paired speaking tests to differentiate the interactional patterns of higher- and lower-proficiency speakers, Galaczi (2014) found that higher-proficiency learners developed topics initiated by other speakers much more than lower-proficiency learners, who preferred to develop their own topics. The author explained this difference in terms of the increased cognitive burden required to develop other-initiated topics, as speaker and listener roles must be kept simultaneously active. In other words, it appears to be too demanding for lower-proficiency learners to decode an interlocutor's utterance at the same time as composing their own (Field, 2011). Therefore, changes in the way learners develop topics could reflect the development of interactional competence.

These results confirmed those of an earlier study, in which Galaczi (2008) identified the following four interactional patterns—collaborative, asymmetric, parallel, and blend—drawing on Storch's (2002) analysis of ESL written discourse. According to Galaczi, a collaborative pattern consists of participants cooperatively extending both their own and their partner's topics, demonstrating alignment in terms of mutuality (i.e., engagement with each other's contributions) and equality (degree of control over the task). This pattern consistently resulted in the highest test scores. A parallel pattern involves participants developing their own topics but not each other's, whereas an asymmetric pattern exists when one participant takes a dominant role in the interaction. Finally, a blend is characterized by two interactional patterns emerging within the same stretch of discourse. In a study carried out at a Turkish university, Artunc and Hart (2019) found that pairing speakers of different proficiency levels together led to asymmetrical patterns of interaction, whereas low-low pairs generated balanced but parallel patterns. Only high-high pairs achieved collaborative patterns, in which the participants extended each other's topics and listened supportively. The results of both studies support Galaczi's (2008) conclusion that learners can benefit from the explicit teaching of the features of collaborative interaction, such as responding to a prior speaker's turn or asking follow-up questions. These are the kinds of skills taught in the course on which the current study is based and are therefore coded and used for analysis. Coding terminology is based on the system used by Galaczi (2014), although adapted for the current context (see Data Coding Procedures in Chapter 3).

Speaking Self-Efficacy

Self-efficacy is the belief of an individual in their ability to successfully undertake an activity (Bandura, 1997). The concept emerged from social cognitive theory and the notion that individuals' beliefs about themselves enable them to manage their own thoughts, feelings, and actions (Bandura, 1986). Self-efficacy is, therefore, chiefly viewed from a cognitive perspective, as opposed to linguistic self-confidence, which is more socially situated and relates to the ability to communicate and identify with the L2 community (Dörnyei & Ryan, 2015). As well as reflecting positive affect towards a task, its importance is underscored by findings that it predicts success to a greater degree than even previous achievements, skills, or knowledge (Schunk, 1991). It is primarily fostered by what Bandura (1997) called mastery experiences, which are previous performances perceived as successful and therefore interpreted positively by the learner. A positive feedback loop can result, as self-efficacy affects people's feelings, behavior, and motivation to continue an activity.

In educational research, it has been found that self-efficacy consistently predicts academic achievement, as students with high self-efficacy tend to expend greater effort, persist in the face of setbacks, and have lower levels of anxiety (Pajares & Schunk, 2001). In terms of language learning, researchers have mainly focused on the relationship between self-efficacy and either general or reading and listening proficiency. For example, Mills et al. (2007) used hierarchical multiple regression to investigate the relationship between self-efficacy and the learning outcomes of over 300 undergraduate learners of French. The authors found that self-efficacy for self-regulation, that is, the belief of students in their ability to plan, monitor, and complete tasks, was the strongest predictor of final course grades ($\beta = .14$), although no causal inferences can be made due to the correlational nature of the analyses. The importance of self-efficacy to language learning achievement was confirmed in Goetze and Driver's (2022) meta-analysis. Based on 37 studies, they established a mean correlation of $r = .48$, which is considered a medium-sized effect (Plonsky & Oswald, 2014), although when subdivided by skill the strongest relationship was with speaking self-efficacy ($r = .56$).

In speaking classes, self-efficacy can be developed by providing learners with repeated opportunities to exchange information and opinions with their peers in a comfortable environment (Mills, 2014). Another advantage of communicative activities is that learners can observe their peers using the target language, thus providing a second, vicarious route to self-efficacy. As Phipps (2024) has pointed out, the 4/3/2 fluency-building activity combines both of these benefits as the frequent change of partner allows numerous opportunities for vicarious experiences, in addition to the potential mastery experiences afforded by the repetition aspect of the activity. Furthermore, observing and interacting with peers can also create what Mills (2014) has referred to as a sense of collective efficacy, or the belief that the group can cooperate to successfully complete similar tasks in future.

An example of how these principles work in practice was provided by Leeming (2017), who investigated how speaking self-efficacy developed among 77 Japanese university students with little previous experience of English oral communication. Working in small groups, communicative objectives were prioritized over accuracy with regular positive feedback also provided. Self-efficacy grew significantly over the course of the academic year, albeit from a low starting point, with questionnaire data backed up by qualitative interviews. Comments such as the following (translated from the original Japanese) reveal the student perspective in such a course: "It was the first time [to have an oral English class] so I was worried at first, but I felt that it was fun and that I would be able to pass. I got used to it and felt that it was fun to speak English." As a counterpoint, Busse and Walter (2013) found that speaking self-efficacy and motivation decreased among British university learners of German, who complained about a lack of oral practice and opportunities to engage with the target language. As well as providing negative evidence for the benefits of communicative teaching methods, this study highlights the fact that increased self-efficacy cannot be taken for granted during a course of study.

In another study based on communicative pedagogy—in this case with pre-vocational Dutch EFL learners—van Batenburg et al. (2019) compared the explicit teaching of interactional strategies with form-focused instruction. The authors reported that the former condition increased learners' "self-confidence" (assumed in this case to be synonymous with self-efficacy) significantly more than form-focused instruction, whether or not the latter was combined with communicative activities. The strategy-based treatment included compensation strategies (e.g., circumlocution), meaning negotiation strategies (e.g., clarification), and audience awareness strategies (e.g., active listening), which the authors noted are commonly absent from EFL textbooks around the world. However, the teaching and practicing of such strategies forms the core of the course on which the current study is based; therefore, whether their use correlates with speaking self-efficacy is investigated.

Language Anxiety

In a general sense, anxiety is "the subjective feeling of tension, apprehension, nervousness, and worry associated with an arousal of the autonomic nervous system" (Spielberger, 1983, p. 15). In terms of language learning, anxiety is seen as situation-specific, of a similar type to test anxiety or stage fright (Horwitz, 2010). Specifically in classroom learning contexts, it has been defined as "a distinct complex of self-perceptions, beliefs, feelings, and behaviors related to classroom language learning arising from the uniqueness of the language learning process (Horwitz et al., 1986, p. 128). Although MacIntyre and Gardner (1989) claimed that foreign language anxiety

(FLA) is unrelated to anxiety as a personality trait, this idea has since been challenged (Dewaele, 2013). For example, Simsek and Dörnyei (2017) found a strong association between general and language anxiety (*r* = .60), indicating that learners with high levels of trait anxiety might be especially susceptible to FLA. Highlighting the complex nature of the construct, Dewaele (2007) has suggested that it lies somewhere between a trait and a state because it is "more sensitive to environmental factors than personality traits and yet more stable than states because it remains relatively stable across [a learner's] languages" (pp. 405–406).

The most frequently used instrument to measure FLA in classroom contexts has traditionally been Horwitz et al.'s (1986) Foreign Language Classroom Anxiety Scale (FLCAS), although skill-specific scales have emerged more recently (MacIntyre & Wang, 2022). Based on such studies, several recent meta-analyses have confirmed that a negative relationship exists between foreign language classroom anxiety and academic achievement. Teimouri et al. (2019) analyzed 97 studies and found a mean correlation of *r* = -.39, whereas Botes et al. (2020) included 66 studies with a mean correlation of *r* = -.36, both of which are considered medium effect sizes. However, despite participants in various studies stating that speaking generates more anxiety than other L2 skills (Baran-Łukarz, 2022), in neither meta-analysis did speaking produce the highest effect size. In Teimouri et al.'s (2019) analysis, listening and writing produced larger effects, whereas Botes et al. (2020) found that oral achievement had the lowest mean effect size (*r* = -.26). In the latter case, the researchers pointed out the heterogeneity of correlations found among speaking studies, which implies that this relationship was influenced by other factors, such as the type of activity being used for assessment. Yet another meta-analysis, conducted by Zhang (2019), found an average correlation between FLA and performance of r = -.34, although in this case speaking anxiety was not reported independently of other skills.

These meta-analyses reveal a clear overall pattern, although to explore the relationship between language anxiety and L2 speaking performance more closely examining individual cases is also informative. In one such study, Phillips (1992) recorded anglophone university learners of L2 French performing two assessment tasks—free speech and a role play. Test scores were correlated with results from the FLCAS, revealing an effect size of r = -.40. This effect is slightly higher than the average reported by Botes et al. (2020), although the context is significant. The test comprised 10% of the course grade and was conducted individually, with the teacher taking part in the role play. Despite attempts to relax the participants beforehand, qualitative analysis revealed how stressful they had found the experience, including one who had broken down in tears causing the test to be delayed for several minutes. This study has since been replicated with Spanish university learners of L2 English (Hewitt & Stephenson, 2012). Similar results were obtained regarding the relationship between an oral test and FLCAS scores (r = -.49). The similarities even extended to witnessing a participant cry during the test, suggesting that such an apparently extreme reaction might not be exceptional in such circumstances. Woodrow (2006) also found a significant negative relationship between L2 speaking anxiety and performance, as measured on an IELTS-type oral assessment, although the major stressor identified by participants was interaction with native speakers, rather than with other learners. Variation among nationality groups also emerged, with Japanese, Korean, and Chinese participants found to be more anxious than their European or Vietnamese peers.

Of particular relevance to the current study, Pan and Zhang (2023) conducted a longitudinal study of FLA and enjoyment during a single semester of communicative teaching with 55 upper-intermediate English majors at a Chinese university. Although enjoyment fluctuated throughout the semester, anxiety remained relatively stable following an initial drop in the first week of the course. The authors speculated that enjoyment is more context dependent than anxiety and therefore more prone to the influence of external factors, such as the behavior of teachers and peers. Inferential statistics were not conducted with the longitudinal data, although the authors did report relatively strong negative correlations between FLA and both self-efficacy and extraversion (r = -.54 and -.55, respectively).

Some researchers have also found a negative relationship between language anxiety and utterance fluency. In a study of low-level learners of Spanish, Pérez Castillejo (2019) found that FLA was a strong predictor of both speed and breakdown fluency under oral exam conditions. The fact that FLA was a much stronger predictor than overall oral proficiency suggests that it can negatively affect cognitive processing to the extent that it overrides linguistic competence. Similarly, Bielak (2022), in a study of mid- to upper-intermediate learners of English, also observed negative correlations between FLA and several speed and breakdown fluency measures, indicating that the effect is not limited to lower-proficiency learners.

In sum, it is clear that speaking in a foreign language can lead to anxiety, although the context in which the speech event takes place is of critical importance. Even within the classroom, some activities are bound to be more stressful than others. For example, a one-to-one examination with a teacher is likely to generate more anxiety than an unassessed group discussion conducted with peers, such as those recorded in the current study. Furthermore, according to social cognitive theory, anxiety—including in academic contexts—is stimulated by a weakened sense of self-efficacy (Bandura, 1997). Conversely, learners with higher levels of self-efficacy tend to exhibit lower levels of anxiety (Pajares & Schunk, 2001), suggesting the existence of a direct negative relationship. Regarding foreign language learning specifically, this relationship was confirmed by Zhou et al.'s (2023) meta-analysis. Although the 37 studies included were not limited to those based on speaking, a strong negative correlation (r = -.70) was found between the two variables. Both speaking self-efficacy and speaking anxiety are measured in the current study.

The Role of Extraversion

It is often assumed that extraverts have an advantage when speaking a second language due to the sociability and talkativeness associated with that personality trait. Furthermore, extraversion has been found to correlate positively with speaking self-efficacy (Leeming, 2017) and negatively with language anxiety (Dewaele, 2013; Pan & Zhang, 2023). On the other hand, extraversion is negatively related to academic success in general, as extraverts tend to be less conscientious and more easily distracted than their introverted peers (O'Connor & Paunonen, 2007). However, it is notoriously difficult to draw firm conclusions about personality and academic achievement. Even in studies that have reported significant correlations between personality factors and learning outcomes, they rarely account for more than 15% of the variance, implying no more than a moderate effect (Dörnyei & Ryan, 2015).

The lack of a linear relationship between personality and achievement can be explained by the myriad contextual factors that interact with personality, such as the learner's age and the teaching methodology used. It makes intuitive sense that some personality types are advantaged by more communicative environments, whereas others might favor studying alone in quieter surroundings, yet many studies do not account for such contextual differences. Indeed, Dewaele and Furnham (1999) suggested that the reason extraversion is held in low regard as an explanatory variable in applied linguistics is that researchers have often failed to distinguish between written and verbal language criteria, thereby neglecting the essential element of context.

When extraversion is correlated with linguistic variables derived solely from speaking activities a clearer pattern emerges, which is that extraverts perform more fluently than introverts, especially in formal or stressful situations (Dewaele & Furnham, 1999). According to Eysenck (1981), introverts suffer from over-arousal of the autonomous nervous system, which can affect parallel processing as working memory becomes overloaded during complex verbal operations. As a result, the speech production of introverts becomes inhibited and utterances become slower and shorter, with an increased number of hesitations (Dewaele & Furnham, 2000). This situation is compounded by the fact that introverts have weaker working memory skills than extraverts to begin with, as measured on short-term recall tasks (Lieberman, 2000). It is also exacerbated during stressful situations, possibly due to introverts expending greater effort to avoid misunderstandings owing to their generally more anxious disposition (Dewaele & Furnham, 2000).

Overall, although results from studies attempting to link personality to L2 learning outcomes have been mixed, this ambiguity is less surprising if we acknowledge that language learning consists of many discrete activities and processes, with some likely to suit extraverts more than introverts and vice versa. For example, Nakatsuhara (2013) found that in a group speaking test, extraverts performed better in open-ended tasks, whereas introverts performed better in more structured activities. This finding suggests that freer spoken interaction, with its potential for heightened stress and demands on working memory, favors extraverted personality types while placing a higher burden on introverts. In addition, Ockey (2011) found that assertiveness, which he defined as one facet of extraversion, was a significant explanatory variable of rated performance in group discussions, especially communication skills and fluency. Ockey's explanation was that assertive participants are more likely to take the lead during interaction and obtain more speaking opportunities, thus also improving their fluency. The relationship between extraversion and speaking performance is also investigated in the current study, which is based on open-ended interaction within a communicative environment.

Gaps in the Literature

Due to its status as a global lingua franca, English as a second language is widely taught around the world. However, there is still much to be learned about how L2 learners develop speaking skills, especially in the case of EFL learners who tend to have limited contact with English speakers and limited opportunities to use the language outside the classroom. This study is therefore designed to address several gaps in the literature.

The first of these gaps relates to the way speaking fluency is measured. Although fluency development has attracted increasing attention in the literature, the focus has largely been on individual learners performing monologue tasks, with the influence of interlocutors removed or minimized. A small number of researchers have attempted to measure L2 fluency using dialogue tasks, with findings indicating that dialogues produce more fluent output than monologues in terms of speed, length of pause, and repair (e.g., Michel, 2011). However, fluency development during group interaction has not been researched, despite the ubiquity of group speaking activities both inside and outside the classroom. Furthermore, the interaction-specific features of fluency have yet to be systematically investigated (Tavakoli, 2016). Therefore, a gap exists in terms of examining fluency development during multi-party interaction, particularly in EFL contexts.

The second gap concerns the quantification of interactional competence in classroom settings. There is a growing body of literature on the development of interactional competence among L2 learners, which reflects the increasing focus on social interaction in SLA. Although this research has its roots in qualitative methods of analysis, there have been some attempts to quantify the features of interaction, such as topic development and listener support moves. However, such studies have focused mainly on high-stakes oral examinations, with coding schemes tailored to the tasks and kinds of language produced in these contexts (e.g., Galaczi, 2014). Interactional competence has not been quantified during unassessed group discussions in EFL classrooms, nor has the effect of teaching formulaic phrases and interactional techniques on the development of interactional competence been investigated.

The third gap concerns the relationship between speaking fluency and interactional competence. Some studies have found moderate correlations between ratings of interactional competence and general speaking proficiency as measured on the TOEFL iBT test (Ockey et al., 2015; Roever & Ikeda, 2021). These findings suggest that the two constructs, although overlapping to some extent, are not completely aligned. However, little is known about the relationship between quantitative measures of interactional competence and, specifically, the fluency of individual learners' speech.

The fourth gap relates to speaking self-efficacy and anxiety in a communicative EFL classroom. In addition to the performance variables of fluency and interactional competence, how students feel about participating in a course based on spoken interaction is also relevant as positive learning experiences are widely believed to have a beneficial effect on learner motivation (Dörnyei & Ushioda, 2011). However, little is understood about how individual difference variables relate to the development of speaking skills. Therefore, the issue of whether affective changes—relating to speaking self-efficacy and anxiety—can be observed during a single semester in a communicative environment is also addressed in this study. The influence of extraversion, as a more stable personality trait, is also examined.

The fifth gap concerns measuring the development of speaking skills in communicative EFL classrooms. Although communicative approaches have been proposed to improve L2 speaking skills for several decades, it is debatable to what extent they have been adopted in EFL environments around the world (Gatbonton &

Segalowitz, 2005; Wei et al., 2018). Furthermore, measuring development in communicative classrooms remains under-researched.

The final gap relates to the impact of pedagogical interventions based on interactive strategies and formulaic language. Several pedagogical frameworks have been suggested for developing L2 fluency, including Nation's (2007) four strands and Gatbonton and Segalowitz's (2005) ACCESS methodology, in which meaning-focused communication is prioritized. Gatbonton and Segalowitz (2005) also proposed the use of prefabricated chunks, given the association of formulaic language with fluency development (Wood, 2010). Another approach to developing speaking skills is to teach interactional strategies, such as paraphrasing or asking for clarification (e.g., "What do you mean?") to deal with communication problems and to facilitate smooth interaction (Dörnyei & Thurrell, 1994). The development of speaking skills during such pedagogical interventions has been under-investigated empirically, although Nakatani (2010) found that use of active listening responses was the most significant predictor of oral test performance among Japanese university students, following a 12-week course of communicative strategy training.

Purposes of the Study

In this study, I aim to investigate students' fluency development during one academic semester in a communicative classroom environment. Data were collected at three points—at the beginning, middle, and end of the semester—allowing progress to be tracked longitudinally. In addition to the monologue data that is traditionally associated with L2 fluency research, data were also collected from group discussions and analyzed in terms of the number and mean length of inter-turn gaps. These measures reflect the co-

constructed nature of fluent interaction and the fact that turn-taking is integral to the smoothness and efficiency of multi-party talk.

A second purpose is to investigate the development of interactional competence during a single semester. The same group discussions—recorded at three points in the semester—are analyzed in terms of topic development and support moves (i.e., the use of interactional micro-skills), and whether the proportion of these moves changed over time. This approach requires the categorizing and quantification of interactional features, which represents what Firth and Wagner (1997) called the etic—or external—perspective of the researcher. However, it is complemented by qualitative analysis, which provides a necessary emic—or insider—perspective, as the next-turn-proof procedure reveals how participants themselves interpret an utterance. In other words, qualitative analysis illuminates how participants orient to the interactional techniques and strategies that form the basis of the pedagogical intervention provided in this study. The data gathered from learner discussions are also compared with data from a discussion recorded by highly proficient speakers of English, to ascertain to what extent they differ.

An additional purpose is to investigate the extent to which the objective measures of utterance fluency and interactional competence correlate with subjective ratings of perceived fluency and communicative adequacy. The issue of whether any development indicated by such objective measures corresponds with the impression made on listeners is examined by comparing them with the subjective judgements of native-speaker raters. Positive relationships would imply that the relevant features contribute to the impression made on listeners, thereby supporting the pedagogical value of fluency development and/or the teaching of interactional techniques and strategies.

A final purpose is to explore any changes in speaking self-efficacy and anxiety in a communicative classroom environment. Communicating directly in an L2 is known to be a face-threatening activity, which is likely to cause anxiety among students unfamiliar with predominantly English-speaking environments (Horwitz et al., 1986). On the other hand, learners with high self-efficacy experience less anxiety (Pajares & Schunk, 2001) and increased motivation (Bandura, 1997). Therefore, if increased speaking self-efficacy and reduced anxiety can be observed during a single semester of study it would provide further support for the effectiveness of such a course. The issue of whether individual difference variables, including the personality trait of extraversion, correlate with the performance variables of fluency and interactional competence is also examined to ascertain whether having a higher degree of one construct offers any advantage regarding another.

Research Questions

The aim of this study is to address the following research questions:

1. To what extent does speaking fluency develop during a one-semester course of spoken interaction?
    - as measured in monologue activities
    - as measured in group discussions
    - as perceived by native-speaker raters
2. To what extent does interactional competence develop during a one-semester course of spoken interaction?
3. To what extent does speaking self-efficacy and anxiety change during a one-semester course of spoken interaction?

4. To what extent is speaking fluency related to interactional competence?

5. To what extent is speaking fluency related to speaking self-efficacy, speaking anxiety, and extraversion?

6. To what extent is interactional competence related to speaking self-efficacy, speaking anxiety, and extraversion?

CHAPTER 3

METHODS

In this chapter, I explain the methods used to answer the six research questions. First, I describe the context of the study and the participants, including ethical considerations. I then describe the instruments and procedures used to gather data, and finally how those data are analyzed.

Context of the Study

The speaking course in this university forms part of the first-year English curriculum. It is a required course, taken twice a week, in both the spring and fall semesters. However, data were only gathered during the spring semester because the intention was to ascertain whether changes could be observed during a single semester. Moreover, students are likely to have different experiences during the summer vacation and their use of English could vary widely, thus undermining the validity of the study if data were gathered over the whole year. Two other required English courses, also held twice a week, are taken in the first year across both semesters: a listening course and a reading and writing course. Although these courses are intended to develop other skills, some opportunities for spoken communication in English are usually provided. Nevertheless, the required speaking course represents the vast majority of the students' spoken English production during the first year in this department.

These classes meet 27 times over 14 weeks for 100 minutes each time. Each class consists of about 20 students, who are expected to use English only in class. Meeting twice a week provides an obvious advantage in terms of speaking opportunities compared with the standard once a week class offered in most departments in most Japanese universities. According to the syllabus, the main objective of the course is to enable students to talk confidently in English about a wide variety of academic and non-academic topics. The development of fluency and communicative ability is also emphasized. Such skills are less likely to have been prioritized prior to entering the university due to the influence of the university entrance examinations. These high-stakes exams dictate that formal English instruction in junior high and high school is primarily geared towards the development of receptive knowledge. The result is that many learners develop uneven skill profiles and enter university with relatively weak speaking abilities due to insufficient practice (Koizumi et al., 2022).

Participants

The participants were first-year Japanese university students in a relatively prestigious, co-educational university in Tokyo. All were members of an interdisciplinary department focusing on liberal arts and social sciences. This department is known for its extensive English curriculum, which is therefore likely to attract students who are receptive to the idea of studying English. Students from my own intact speaking classes, consisting of about 20 in each class, were sought to participate in the study. Although more were recorded, 20 core participants were selected for analysis, to allow for attrition or non-attendance in one or more of the recorded or data collection sessions. Only students who completed all of the recorded activities and questionnaires were considered. In addition, selection aimed to reflect a range of speaking ability and individual difference factors (i.e., extraversion, self-efficacy, and anxiety), based on the recorded activities and questionnaires completed during the study. An equal gender balance was also targeted, although only 9 male participants met the criteria for selection; therefore, the final list of core participants comprised 11 female and 9 male participants.

Prior to enrolment, all students had completed a minimum of six years of formal English study in junior and high school. First-year students in this department are placed in their required English classes based on their scores on a standardized test, the TOEFL ITP, which contains listening, reading, and writing elements. The test is taken in April, prior to the first semester, with students then divided into three levels. According to the university, these levels approximately correspond to CEFR levels B2 (upper-intermediate), B1 (intermediate), and A2 (low-intermediate), with textbooks and teaching materials selected on this basis. Cut-off points vary depending on each year's intake, and are not disclosed by the university, although the average score obtained in the year this study was undertaken was 496.52. All student participants in the current study were drawn from two Level 3 classes, which is the lowest level in this department. As the sample size is relatively small, using participants from the same level enhances the generalizability of the study, albeit primarily to learners of a similar level. However, the fact that the placement test does not contain a speaking component means that classes are often mixed in terms of their speaking proficiency.

Although a fully experimental design—with learners drawn at random from the population of interest—would enhance generalizability, the fact that it was conducted in a classroom setting adds pedagogical authenticity to the study (DeKeyser & Botana, 2019). The participants were divided into groups of three for the recorded group discussions, as they often were during discussion activities conducted in the regular lessons. Groups consisting of more or less than this number were not used for analysis. Non-core participants were not analyzed individually, although their participation in the group discussions means that their behavior influenced the data taken from that context.

Informed consent was sought prior to the study and participants were notified that this consent and cooperation could withdrawn at any time (see Appendix A for the English version and Appendix B for the Japanese version—the Japanese version was used). Furthermore, the participants were informed that materials collected would not be used for purposes other than this study, such as promotional or any other purposes of the university, and they were given the right to request that their data not be used. As the participants were drawn from my own regular classes, it was ethically imperative that all students in these classes knew that they would not be penalized in any way if they did not agree to participate in the study. Permission to conduct the study was also obtained in advance from the head of the relevant university department. This was sufficient for IRB (Institutional Review Board) purposes at that university, although a letter confirming this permission was sent to the Temple University IRB office as part of that application process.

In addition to the student participants, three native-speaker English teachers were recruited to rate samples from the recorded group discussions. The raters were required to undergo online CITI (Collaborative Institutional Training Initiative) training in human research before watching the video recordings. I completed the same online training course prior to data collection. Three highly proficient Japanese speakers of English were also sought to record a single group discussion for comparison purposes. They were recruited from Temple University Japan's postgraduate courses in education or applied linguistics, in which a minimum TOEFL score of 90 (Internet-based test) or 575 (paper-based test) is required for matriculation.

Instrumentation

In this section I introduce the instruments and materials used—both for pedagogical and analysis purposes—in this study. The main pedagogical instruments were the textbook and the supplementary discussion-based activities provided by me. Analysis was based on video-recorded activities, which were recorded during class time, as well as the speaking self-efficacy, anxiety, and extraversion questionnaires.

Background Questionnaire

A background questionnaire was administered at the beginning of the course, to ascertain the students' previous experiences of learning English. It included formal test scores (e.g., TOEIC and TOEFL), overseas experience, and English study experience both in and out of class (see Appendix C for the English version and Appendix D for the Japanese version). The students were given the Japanese version of the questionnaire to complete during the first week of the course. The purpose of this questionnaire was to shed light on students' performance during the recorded activities. For example, it might be the case that higher-performing students had more study-abroad experience or had studied English extensively in cram schools. Also, if some students had previously experienced communicative English lessons they might be better prepared for this kind of communicative course. It could also help to explain discrepancies between linguistic proficiency, as evidenced on the placement test, and speaking proficiency.

The Course Schedule

The course was based on a variety of topics related to university life and the wider society, provided both by myself (as the instructor) and a textbook assigned by the university. The first lesson of each week was a discussion-based lesson, with the topics and activities chosen by myself, based on my previous experience of teaching such classes (see Table 1—textbook and non-discussion-related topics are in italics). Several discussion skills (also shown in Table 1) were introduced during the course, focusing on both speaker and listener roles. These skills were presented either via formulaic phrases (e.g., for giving and supporting opinions) or interactional techniques (e.g., asking follow-up questions). Students were then encouraged to use these phrases and techniques during pair and, especially, group discussions (see Appendix E for a full list of the formulaic phrases and questions introduced).

The order in which these skills were introduced was arbitrary, although the intention was to present the most fundamental ones earlier in the course. For example, as the heart of any extended interaction is the content generated by the participants, giving opinions and reasons both appeared very early. Once the ability to produce an extended turn had been established, organizational skills (Joining a Discussion and then Choosing Topics) were introduced. Also, some foundational interactional techniques (agreeing and disagreeing, asking follow-up questions, using response tokens such as "Uh-huh" and "I see", and checking understanding) had already been briefly introduced in Lesson 1 (Course Introduction), so if they appeared later in the course the purpose was to reinforce and expand on their earlier introduction. Paraphrasing appeared last because reformulating a comment is a relatively complex linguistic skill. It is also arguably less fundamental than some of the skills introduced earlier because it is highly contingent on the interlocutor and therefore not always necessary.

Table 1. *Course Schedule*

Week	Lesson topic	Discussion skill
1	Course Introduction *Questionnaire **	
2	The Importance of Communication *Family Connections*	Giving Opinions
3	Making Friends at University *Music*	Giving Reasons
4	Going to University *Good Design*	Agreeing and Disagreeing
5	University Entrance Exams *Inspiration*	Joining a Discussion
6	Environment *Presentation Preparation **	Giving Examples
7	Country vs. City Life *Presentation 1*	Review of Skills
8	Discussion Test 1 *Ethical Choices*	——
9	Becoming Independent *Giving*	Follow-up Questions
10	Social Pressure *Nature*	Talking about Possibilities
11	Face-to-Face vs. Online Communication *Better Cities*	Choosing Topics
12	Technology *Presentation Preparation **	Paraphrasing
13	Learning Values *Presentation 2*	Review of Skills
14	Discussion Test 2	——

Note. * = data collection point.

Students were assessed on their class participation (20%), homework reading and other assignments (20%), two presentations (20%), and two discussion tests (40%)— the latter two held in the middle and at the end of the semester. Participation was expected to be in English only and included the use of the target discussion skills. In regular discussion lessons, I monitored and provided feedback on students' use of the skills and formulaic language introduced. In discussion test lessons, I observed individual groups (of three or four members) one at a time during an extended (10- to 15-minute) discussion. A numerical score for each participant was generated based on the following five categories: speaking turns, questions, response tokens (including reactions, agreement tokens, and clarification moves), discussion skill use, and avoiding L1 use or incomprehensible turns. All turns and interactional moves were tallied, up to a maximum of five in each category, giving a raw score out of 25. This raw score was then converted to a score out of 20 for grading purposes. See Appendix F for an example discussion test scoresheet.

Lesson Activities

Meaning-focused output was heavily prioritized in the discussion-based lessons. A typical lesson was centered around a single content theme and consisted of the following activities. Prior to the lesson, the students read an 800-1,000-word, level-specific text for homework, from which they were required to note three interesting ideas or facts. These lessons began with a homework check, followed by a warm-up speaking activity, conducted in pairs, related to that lesson's theme. For example, in Week 7, the discussion-based lesson's theme was Country versus City Life. The warm-up questions, displayed on the board or a screen, were as follows:

1. What do you like and dislike about your hometown?
2. Where would you like to live in the future?

Following the warm-up, a discussion skill—with any associated formulaic language—was presented and practiced in a semi-controlled manner in pairs or small groups. For example, the Week 7 discussion skill was Agreeing and Disagreeing (see

Appendix G for the lesson handout). In this lesson, after the skill was presented, the students discussed the Practice questions (page 1 of the handout) in pairs. The questions in this semi-controlled activity were not always directly related to the lesson theme, as the purpose was primarily for the students to practice using the formulaic language. I monitored and ensured that the skill and associated language was being used appropriately. Feedback was given on the spot or at the end of the activity as necessary.

The two group discussions were preceded by preparation activities (see Appendix G, on pages 2 and 3 of the handout). These preparation activities were designed as pair work speaking activities, although one was usually conducted in the form of a repeated monologue activity to further encourage fluency development (Nation, 1989). Discussion preparation activities were geared towards providing ideas and generating content for the extended group discussions, which consisted of either three or four members (or five in exceptional circumstances) and lasted from 10 to 15 minutes at my discretion. Either two or three question prompts were provided for each group discussion (see Appendix G, at the bottom of pages 2 and 3), although students were made aware that it was their responsibility to sustain the interaction. In principle, I did not intervene during group discussions and feedback was generally given after the activities had finished, so as not to interrupt the flow of communication.

Throughout the lesson, feedback was typically centered on the students' use of the discussion skills, especially the one introduced in that lesson. Such feedback could be given during any stage of the lesson and often consisted of pointing out effective use of a target phrase or question, thus positively reinforcing its use. It could also consist of pointing out when a target form would have been appropriate but was not produced, as it is during meaning-focused communication that the need for linguistic elements is at its most salient, and thus memorable, for the learner (Long, 1998).

However, in keeping with the student-centered ethos of the course, feedback did not have to be teacher-fronted. Alternatively, students—either individually or in pairs or small groups—could reflect on their use of the target skills and language, as well as how they could improve in future activities. Students could also be asked to reflect on the content of a group discussion, for instance by raising their hands in answer to one of the question prompts. Table 2 shows the lesson flow of a typical discussion-based lesson.

Table 2. *Regular Discussion Lesson Flow*

Time	Activity
5 mins	Homework check
5 mins	Warm-up activity (pair)
10 mins	Skill presentation
10 mins	Semi-controlled practice (pair)
10 mins	Discussion 1 preparation (pair)
15 mins	Discussion 1 (group)
5 mins	*Break*
15 mins	Discussion 2 preparation (monologue)
20 mins	Discussion 2 (group)
5 mins	Wrap-up

Review lessons differed from regular lessons only in that previously introduced skills were reviewed and practiced further, rather than a new skill being presented. During test lessons, I listened to each group one at a time, while the other groups waited outside, and tallied the contributions of each participant. The first lesson was also distinct in that the aims and expectations of the course were laid out and four basic interactional techniques were introduced and briefly practiced in this lesson: agreeing and disagreeing, asking follow-up questions, using response tokens, and checking understanding. It was made clear to students that they were expected to use these techniques in every lesson when appropriate. Students also performed a repeated monologue activity and a 10-minute group discussion, both of which were likely to be unfamiliar activities.

Textbook lessons were always the second lesson in the week and differed from the discussion-based lessons in that they covered a wider range of skills. A typical textbook lesson included short reading or listening, writing, grammar, and vocabulary exercises—in addition to speaking and presentation activities—all related to that unit's theme (see Table 1). A repeated monologue activity was also included in the textbook lessons, based on the lesson theme. Pedagogically, these lessons contributed to all elements of Nation's (2007) four strands framework, although speaking was prioritized. In sum, a large amount of meaning-focused output—recommended in several pedagogical frameworks that aim to develop L2 fluency (e.g., Gatbonton & Segalowitz, 2005; Nation, 2007)—was provided throughout this twice-weekly course. Fluency development was further promoted by the regular use of repeated monologue activities appropriate to the learners' proficiency level.

Recorded Activities

The recording sessions began with a two-minute monologue delivered by each participant (see Appendix H for an example from Time 3), followed by a group discussion (see Appendix I for an example excerpt from Time 3). Although these activities were unfamiliar to the participants prior to the course, the fact that monologue and group discussion activities had been conducted in Lesson 1 meant they had experienced them prior to the first recording session. The total duration of each recorded group discussion was nine minutes, so that in theory each member had approximately three minutes' speaking time. Each group consisted of members from the same class, although the groups were mixed so that each group consisted of different members during at each recording point.

There are several advantages to mixing the group members in this way. First, it minimizes the problems of attrition and non-attendance that could be caused by attempting to keep all members consistent. Second, it can be observed whether changes in performance occur independently of the interlocutors. In other words, if all groups consisted of the same members each time it could be argued that any changes over time were contingent on that particular group dynamic and might not have occurred with different interlocutors. Third, the students were used to this procedure from regular in-class group discussions. Finally, it mitigates the possibility that improved performance is due to increased familiarity with each other. Communicative language learning takes place socially, so personal relationships are bound to affect this process and the development of these relationships is likely to influence the development of interactional competence (Taguchi, 2014). For this reason, participants who often work together in regular lessons were not placed in the same recording groups, which always consisted of at least one male and one female participant. In regular lessons, male and female students rarely sat together so they were less likely to be familiar with talking to each other. Another benefit of this approach is that it is easier to distinguish between voices when listening to the recordings.

The main purpose of recording the monologues was to observe whether the participants' utterance fluency developed during the course. A secondary purpose was to investigate whether any changes that occurred in the group context—in either fluency or interactional competence—corresponded to changes in utterance fluency, as measured during the monologues. Conversely, little or no change in performance during the monologue activities over time would imply that any changes that occurred in the group context were likely to be unrelated to the underlying fluency of the individual. Pedagogically, the monologue activities also functioned as a warm-up for the group discussions. Ideas and content generated by the participants for the monologue could be recycled during the group discussions, in the same way that discussion preparation activities served to generate content for group discussions during regular lessons. The same procedure was strictly adhered to in all three recordings.

The three topics chosen for the recording sessions were: seasons, places to visit in Japan, and high school life versus university life (see Appendix J), which were selected on the basis that they should be familiar enough for first-year university students to talk about with relative ease. Regarding fluency, the aim was to measure the participants' output as accurately as possible. Conceptually difficult topics, requiring more complex ideas, were therefore avoided so as not to distort this measure due to increased cognitive load (Skehan, 2009). Furthermore, it was considered unlikely that these topics would require sophisticated vocabulary, which could hinder fluent performance for students at this proficiency level (Nation, 2013).

Another important consideration was that these topics were different from those used for group discussions during the regular lessons. If previous topics were used, even several weeks apart, fluency gains could arguably be explained by the priming effects of task repetition (Bygate, 2001). On the other hand, if such gains are observed across novel topics, this provides greater evidence that the processes underlying speech production have been affected (De Jong & Perfetti, 2011).

Speaking Self-Efficacy and Anxiety Questionnaires

The speaking self-efficacy and anxiety questionnaires were administered before each recording session, to ascertain any changes in reported self-efficacy or anxiety (see Appendices K and M for the English versions and Appendices L and N for the Japanese versions—the Japanese versions were used). The self-efficacy items were adapted from a questionnaire that has previously been used with Japanese university students, demonstrating acceptable validity in that context (Leeming, 2017). All items were specific to the learning context, as recommended by Bandura (2006) in the case of self-efficacy, and Zhang (2019) in the case of anxiety. The participants were asked to indicate how confident, or anxious, they felt about speaking English in a variety of situations related to the speaking class they took. A six-point Likert scale was used, with participants selecting *Strongly disagree*, *Disagree*, *Moderately disagree*, *Moderately agree*, *Agree*, or *Strongly agree* in response to a series of statements. An even number was chosen to avoid a middle option that might be relied on too heavily by respondents (Dörnyei, 2010). Whether there is any relation between speaking self-efficacy/anxiety and fluency or interactional competence could then be empirically investigated via correlation analysis. The participants were also asked to answer open-ended questions after the final recording to investigate what aspects of the course they felt were most useful and in what ways, if any, they felt they had improved the most. Comments from this section were used to supplement the empirical findings mentioned above.

In terms of a hierarchy of difficulty, I hypothesized that the self-efficacy items 1, *I can express my opinion in English*, and 3, *I can speak English when talking to a partner in pair work activities*, would be the easiest for participants to endorse. My reasons were that giving an opinion is a fundamental aspect of communication and generalizable to all discussion activities, whereas talking to a partner is less face-threatening than talking in a group. In contrast, I expected that items 6, *I can solve communication problems in English*, and 8, *I can speak English only for a whole lesson*, would be the most difficult to endorse. My rationale was that the students would not be used to speaking English for such a long time, whereas solving problems is contingent on what happens in the discussion and implies that something has gone wrong and needs to be fixed.

Regarding the anxiety section, I hypothesized that items 13, *I am worried about making mistakes when speaking English*, and 14, *I feel nervous when someone doesn't understand what I say in English*, would be more likely to be endorsed (i.e., agreed with). In the former case, this is because accuracy tends to be heavily prioritized in pre-university formal English learning, whereas the latter case is an example of the communication problems mentioned above, and therefore likely to induce anxiety. On the other hand, I expected that items 9, *I feel nervous when speaking to a partner in pair work activities*, and 11, *I feel nervous because other students speak English better than I do*, would be less likely to be endorsed. My rationale was that, as mentioned previously, pair work activities are less likely to be anxiety-inducing than group activities, whereas in the latter case students would be aware that they had all been placed in the same proficiency level.

Extraversion Questionnaire

The extraversion questionnaire was administered once, as dispositional personality traits are considered more stable than contextualized characteristics such as self-efficacy (McAdams & Pals, 2006), and are therefore less likely to change during a single-semester course of study. The items in this questionnaire were directly taken from one that had previously been used with Japanese university students and demonstrated acceptable validity (Apple, 2011). As with the speaking self-efficacy questionnaire, the participants indicated whether they agreed with a series of statements, with the same six options ranging from *Strongly disagree* to *Strongly agree* provided as above (see Appendix O for the English version and Appendix P for the Japanese version—the Japanese version was used).

Regarding the hierarchy of difficulty, I hypothesized that items 3, *I start conversations*, and 6, *I talk a lot*, would be the easiest to endorse because they do not specify the communicative context. For example, talking a lot could apply equally to being at home and being outside with a large group of people. In contrast, I expected that items 1, *I am the life of the party*, and 4, *I can get along with almost anyone*, would be the most difficult to endorse. In the former case, by definition not many people can fulfill this role, whereas the latter case implies an ability to interact smoothly with people outside one's own family or social circle, even complete strangers. I expected the other items (2, 5, 7, and 8) to be moderately difficult to endorse because they either imply an active role but not necessarily in a large group (for example, *I am skilled at dealing with people*), or being in a large group but not necessarily taking an active role (for example, *I feel at ease in a large group of people*).

End of Course Questionnaire

The end of course questionnaire was issued after the final recording session. It consisted of only four questions (see Appendix Q) and participants were instructed that they could answer in either English or Japanese. Questions were open ended and allowed participants to state which activities and formulaic phrases they found most useful, as well as in what way or ways they improved, during the discussion course. This questionnaire was not obligatory and was conducted anonymously to encourage the participants to answer frankly. Although the answers could not be attributed to any individual, it was hoped that comments would provide an emic perspective and illuminate the participants' feelings about the course in general. Such comments could then be used to supplement findings from the other forms of data, whether performance- or questionnaire-related.

Highly Proficient Speaker Comparison

To provide a point of comparison, data were also gathered from three highly proficient Japanese speakers of English. It must be emphasized, however, that the aim of this comparison was not to highlight deficiencies in the lower-proficiency learners' speech production, but rather to facilitate a deeper understanding of the constructs of interest in this study. As Japanese speakers of English, they also represented a more realistic target for learners who are still developing their own language skills. Furthermore, unplanned speech is often challenging even for highly proficient speakers, and apparent disfluencies—such as hesitations and repetitions—are not only commonplace but can facilitate successful communication (De Jong, 2016). Therefore, if such features are observed among highly proficient speakers, it forces researchers to challenge the notion of what constitutes proficient or competent speech. Comparing low-intermediate learner and highly proficient speaker data also allows assumptions regarding what constitutes the development of interactional competence to be tested.

A group of three highly proficient speakers were recorded discussing the topic about favorite places to visit in Japan (see Appendix J). This topic was chosen as there are a wide range of possible answers, which might stimulate such speakers more than talking about seasons. The high school versus university topic was avoided because, as postgraduate students, it would require these participants to talk about their experiences in a different tense than freshman undergraduates.

Research Design

Quantitative analysis was used to answer the research questions relating to fluency and individual differences—speaking self-efficacy, speaking anxiety, and extraversion—as numerical data were produced. These data were either in the form of temporal and utterance fluency measures, or Likert scale rating and questionnaire responses. However, to answer the second research question—regarding the extent to which interactional competence develops during a one-semester course of spoken interaction—a convergent mixed methods research design was employed (see Figure 3).

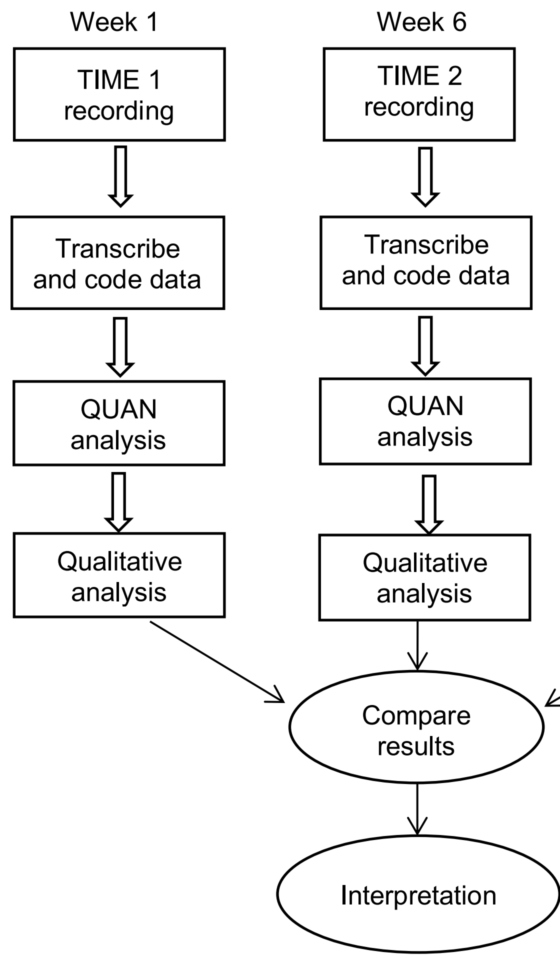

Figure 3. The Convergent Research Design

Mixed methods research involves collecting and integrating both quantitative and qualitative data, based on the assumption that combining research paradigms provides a deeper understanding of the phenomena in question than relying on either approach alone (Creswell & Creswell, 2018). Convergent designs are typically used when two separate sets of data—one quantitative and one qualitative—are collected and then merged in order to compare or combine the results. Such designs are popular in mixed methods research because they are efficient and allow both types of data to be collected during the same phase of the research, rather than sequentially. Convergent designs are also useful for corroborating or illuminating quantitative results with qualitative findings (or vice versa). However, it can be challenging to merge different types of data (e.g., numerical and textual) and resolve contradictions between them (Creswell & Plano Clark, 2018). In the current study, both quantitative and qualitative data sets came from the same source—recorded group discussions—which were transcribed and coded based on the features of interaction. These features were then counted, thus transforming the transcriptions into numerical data, which were then compared with the qualitative data (see Analyses, below, for a fuller explanation).

The rationale for using a convergent design in the current study was that qualitative methods could complement the interactional quantitative data by providing some context and revealing how the participants themselves had interpreted an utterance. For example, although tallying the use of specific interactional moves (e.g., questions or response tokens) at different points in the course allows development to be tracked longitudinally, it does not reveal the context in which they are used. Therefore, qualitative analysis was used to confirm whether participants had used these interactional features in a pragmatically appropriate way. Due to the relatively small number of participants and the lack of a comparison group, no causal claims can be made regarding any such development; therefore, it is a descriptive, rather than an experimental study. However, having a limited number of participants allows the issue to be approached from multiple perspectives, which has been called for by many researchers in the field (e.g., Galaczi, 2014; Segalowitz, 2010). It also allows a deeper analysis of their output to be conducted, given the limited time available to an individual researcher.

Three criteria need to be considered when conducting a mixed methods study: the timing, weighting, and mixing of the quantitative and qualitative elements (Creswell & Plano Clark, 2007). Regarding timing, although the same data source was used for both types of analysis, the quantitative analysis preceded the qualitative analysis because the latter was used initially for confirmatory purposes. The former also has greater weighting within the design as not only is it used to answer more of the research questions, but it was also used to draw provisional conclusions regarding the development of interactional competence. Regarding mixing, in a convergent mixed methods design quantitative and qualitative data are typically merged when interpreting the results (Creswell & Creswell, 2018). Accordingly, following the analysis of the qualitative data the two types of analysis are integrated during the discussion section, when the provisional conclusions drawn from the quantitative results are interpreted in light of the qualitative findings.

Procedures

In this section, the procedures for collecting and recording data are described. Video and audio recordings were made of nine-minute group discussions, preceded by two-minute (audio only) monologues made by each participant. These recordings were made during class time in the regular classroom. At the beginning of those lessons, the students were assigned to a recording group and given specific times to return to the classroom. Two groups were recorded simultaneously, which means that between six students were in the classroom at one time, unless there was an odd number of groups in which case one group was recorded separately. Audio recordings were used for a homework reflection activity (described below).

Participants first recorded the monologue, having been given the topic on a piece of paper. They were instructed to say as much as they could—no guidelines were given regarding formulaic language or skill use because this activity was for fluency-measuring purposes (see Appendix J for the topics and instructions). One minute preparation time was allowed. The participants were each given an IC recorder and, after being shown how to use it, dispersed around the classroom to record their monologues, which they began by stating their name for voice identification purposes. All monologues were recorded without an interlocutor present so that an extended stretch of speech could be produced without interference or assistance (Tavakoli & Wright, 2020).

After completing the monologues and saving their output, the group discussions were then recorded. The two groups sat at opposite ends of the classroom and were recorded on separate video recorders. The classroom was large enough for each group not to be distracted by the other. IC recorders, placed on a small desk in front of each group, were also used for audio-only recordings. The topic was the same as for the monologue and the participants were given the discussion questions on a new piece of paper (see Appendix J). It was imperative for the group discussions to reflect the ability of the learners to speak and interact independently; therefore, I did not intervene once the participants had confirmed their understanding of the task instructions and the recording had begun. The participants, too, were verbally instructed not to ask for assistance and were made aware that the tasks were to be carried out with their peers only.

The written instructions, which were provided in English and Japanese (see Appendix J) made it clear that the participants were to answer the questions in the way that they felt was best. This section was underlined so that the participants did not feel under any obligation to use the formulaic phrases or skills that they had learned in class. I also emphasized that this activity was not a test and would not affect their grade in any way. Therefore, if any such skills or phrases were used it should have been because the participants found them useful, rather than because of any extrinsic desire to please me, as the instructor. Such desirability effects cannot be ruled out entirely, but the intention was to ascertain which skills and phrases the participants valued for their efficacy in a group discussion. Encouraging autonomy in this way helps to provide an emic—or participant-relevant—perspective, given that only the output of learning, rather than the internal processes, are available for analysis (Hall & Pekarek Doehler, 2011). Exemplars of the output produced were used to highlight and examine the participants' use of the discussion skills and formulaic language, with all group discussions manually transcribed and coded.

Each group discussion was nine minutes in duration, timed from the moment the first participant said anything relevant to the topic, such as reading the first question out. If students greeted each other, for example, or stated their names, this did not count as starting the discussion. The recording devices continued for at least 30 seconds after the official time elapsed so that any comment that began before the time was up could still be included in the analysis. No pre-task planning time was afforded, besides the monologue activities, as authentic spoken communication outside the classroom rarely allows for planning time. Moreover, pre-task planning can be counter-productive, resulting in long, disconnected turns as speakers focus on what they have planned to say rather than responding to or developing their partner's topic (Nitta & Nakatsuhara, 2014). Online planning could, of course, take place during the interaction.

It is impossible to ensure that all topics are of precisely the same level of difficulty and familiarity to all participants. However, to reduce the bias introduced by using multiple topics, they were counterbalanced between the two classes (see Table 3). An equal number of participants were selected for analysis from each class (ten from each) to reflect this counterbalancing.

Table 3. *Schedule for Recorded Activities*

Class	Session 1	Session 2	Session 3
A	Topic 1	Topic 3	Topic 2
B	Topic 2	Topic 1	Topic 3

The recording sessions took place at the beginning, middle, and end of the course, in Weeks 1, 6, and 12 (see Table 1). The first recording was made in the second lesson of Week 1, as the anxiety caused by recording in the first lesson could have resulted in artificially low levels of performance. Moreover, without experiencing an orientation lesson in which the monologue and group discussion activities were introduced and practiced, students might not have fully understood how to participate in the recorded activities, thus undermining performance in the first recording session.

The recordings were made during textbook rather than discussion-based lessons, to reduce the risk of priming effects related to the discussion skills and strategies. For example, if the students had just studied and practiced a particular skill or set of formulaic phrases, it is likely that those features would appear prominently in a recording made during that lesson, thus artificially inflating their use. Nevertheless, the order in which the discussion skills were taught is bound to influence the recorded data. For example, students would be more likely to use paraphrasing techniques in recording

Session 3 than Session 1 as it was taught in the previous lesson, at the beginning of Week 12. However, this merely reflects the way that the course is taught—one would expect it to influence what the students learn, and therefore subsequently produce.

As mentioned previously, the students recorded at different times in the lesson, with two groups recording simultaneously. The students who were not recording at that time were given other activities to work on in a different room. During recording Session 1, they were asked to complete the background questionnaire, while during Sessions 2 and 3 they were required to prepare presentations (related to textbook topics) in a computer room. In addition, all students were required to listen to their group discussion and complete a reflection task for homework. The audio of the discussion was sent to each student individually and they noted down two or three things they thought they had done well and that they needed to improve (e.g., express opinions, ask questions, use reactions). The purpose of this homework activity was to encourage self-reflection and to allow the students to track their own progress throughout the course.

The speaking self-efficacy and anxiety questionnaire was administered three times, in the lesson before each recording session (due to time restrictions in the recording lesson). The extraversion questionnaire was administered once, in the lesson before the first recording session. The end-of-course questionnaire was administered once, in the lesson after the final recording session. All questionnaires were conducted at the beginning of those lessons so that the students' responses were not influenced by that lesson's activities. For example, if they had been distributed after a pair work speaking activity, the responses regarding pair work activities could have been affected by how each student felt during that activity. All students were asked to complete the questionnaires, as they were distributed during class time. A space was provided for the students to write their name, although they were informed that doing so was optional.

Data Coding Procedures

Monologue Fluency

The recorded data from the 2-minute monologues were transcribed in Microsoft Word and analyzed manually, with pause duration measured using Praat (Boersma & Weenink, 2021) speech analysis software. The first 90 seconds of the monologues were used for analysis, beginning from the time the participant began speaking. As participants began speaking at slightly different times, using the full two minutes would have produced speech samples of differing lengths, which would obscure the utterance fluency measures. On the other hand, using less than 90 seconds might not have provided enough material to separate the participants in terms of their performance. For example, the one-minute planning time might provide enough material to speak from memory for the first part of the monologue, but speakers were more likely to have to rely on online planning towards the end. Online planning is more closely associated with the kind of mid-clause pauses and disfluencies that are indicative of learner speech (Tavakoli, 2011); therefore, sufficient time was allowed for this process to be reflected in the analyzed samples.

For measurement purposes, utterance fluency is commonly subdivided into three components: speed of delivery, breakdown fluency, and repair fluency (Tavakoli & Skehan, 2005). Breakdown fluency relates to the number, duration, and placement of pauses—whether silent or containing vocalized fillers—made during speech. All number of pause measures were calculated per minute of speech, whereas duration of silent pause measures were calculated as mean values. The assumption is that fewer and shorter pauses represent more fluent output, especially when such pauses occur mid-clause.

Regarding silent pauses, it must be decided how long a non-verbal gap is before it is counted as a pause; in other words, at what point a pause interrupts the flow of fluent speech. This cut-off point has varied among researchers, usually between 200 to 400ms, although De Jong and Bosker (2013) concluded that 250ms is the optimum number for L2 research because it leads to the highest correlation between number of silent pauses and proficiency as indexed by vocabulary knowledge. Moreover, they argued that cut-off points above 300ms confound the number and duration of pauses because if only long pauses are counted, this number is bound to be strongly related to the mean duration of pauses. Accordingly, 250ms is the cut-off point adopted in this study.

Some researchers have also capped silent pause length in an attempt to minimize the influence of the environment. For example, De Jong and Perfetti's 2011 set an upper limit of 2.5 standard deviations above the mean length of pause in a specific monologue to eliminate pauses that might have resulted from external distractions rather than a student's fluency. However, following Doe (2025), pause length was not adjusted in the current study as my impression was that even relatively long pauses were caused by students attempting to generate content for their speeches rather than by any task-irrelevant distractions. Furthermore, students who stopped speaking too early, and thus did not complete a monologue task, were not considered for selection as a core participant.

Having established what constitutes a silent pause, it is important to define what is meant by a filled pause. Such pauses commonly consist of non-lexical vocalizations such as *er* and *um*, although I have also counted lexical fillers (e.g., *well* and *you know*) as filled pauses if used as stalling devices (Peltonen, 2017). Sound stretches, such as an elongated /i/ sound in the word *the*, are also counted as filled pauses because they are considered to serve the same purpose (Freed, 2000).

Another element of breakdown fluency is the location of a speaker's pauses. Pauses at clause or syntactic boundaries are a natural part of fluent speech and are used by L1 and L2 speakers alike for conceptual planning of the upcoming message. Pauses that occur mid-clause, by contrast, are likely to be caused by formulating the message mid-utterance (Skehan et al. 2016). As a result, such pauses are more disruptive and are closely associated with the L2-specific cognitive processes reflected in learner speech (Kahng, 2020). Therefore, pause location was analyzed as a further indicator of breakdown fluency.

Speed of delivery refers to how fast a sample of speech is articulated; in other words, the density of language produced per unit of time (Tavakoli & Skehan, 2005). However, using overall speech rate as a measure —which includes pauses—is confounded with breakdown fluency, as the more a speaker pauses the slower the speech rate becomes. On the other hand, articulation rate—measured by dividing the total number of syllables by total phonation time excluding pauses—produces a statistic unrelated to any other measure of fluency and is, therefore, a purer measure of speaking speed. Following De Jong et al. (2013), I use mean syllable duration, which is the inverse measure of articulation rate (i.e., dividing the total phonation time by the number of syllables). The benefit of using this inverse measure is that, in line with other measures of fluency, a higher value (i.e., longer mean syllable time) represents less fluent speech.

Syllables were counted by pasting the transcripts into an online syllable counter (http://www.syllablecount.com/). Only Japanese proper nouns, such as place names, were included in the analyses and were checked and adjusted if necessary. Each Japanese mora was counted as one syllable.

Finally, repair fluency is concerned with the number of repetitions and self-corrections made by a speaker. Both measures were calculated per minute of speech. Repetitions can be of words, parts of words, or whole phrases. Self-correction is the reformulation of an utterance, or part of an utterance, whether occurring at the beginning (i.e., a false start) or later in a speech sample. Repair fluency reflects a distinct aspect of cognitive processing from breakdown fluency because it is concerned with the online modification of speech (Tavakoli & Skehan, 2005). In other words, it reflects the monitoring (rather than the formulation) stage of speech production. Table 4 presents all utterance fluency measures used in this study.

An overall measure of pruned speech rate is also included, which encompasses all three elements of speed, breakdown, and repair (De Jong, 2016). Not only does this statistic provide a single measure of utterance fluency for comparison purposes, but it can also be used for correlation analysis with the interactional competence and individual difference variables (see the Analyses section, below). Pruned speech rate is calculated by removing all disfluent syllables from a speech sample and dividing the remaining number of syllables by the total time. It therefore incorporates speed and breakdown—as pauses are not removed—and repair information, as repetitions and reformulations are reflected in the time taken to produce the sample.

Table 4. *Monologue Utterance Fluency Measures*

Category	Measure
Speed	Mean syllable duration
Breakdown	Number of silent pauses (end-clause)
	Number of silent pauses (mid-clause)
	Number of filled pauses (end-clause)
	Number of filled pauses (mid-clause)
	Mean duration of silent pauses (end-clause)
	Mean duration of silent pauses (mid-clause)
Repair	Number of repetitions
	Number of repairs
Overall	Pruned speech rate

Note. Number of pauses, repetitions, repairs, and speech rate are calculated per minute.

Measuring Fluency in Group Discussions

The many possible interruptions, efforts to assist, or overlaps caused by multiple interlocutors mean that an individual's utterance fluency cannot be measured reliably during group discussions. Furthermore, pauses, disfluencies, and variable articulation speed can serve communicative functions during interaction (Pallotti, 2020), meaning that counts of such features might not reflect an individual's underlying cognitive fluency. Fluency is a broad concept, however, and there are aspects that can be usefully measured during multi-party interaction (see Table 5), including the amount of speech produced by each participant. More speech indexes greater participation, which in turn could reflect heightened self-efficacy. In this study, the amount of speech produced during group discussions was calculated by tallying the total number of fluent syllables (i.e., with filled pauses, repeated syllables, and false starts removed) generated by each participant per minute. This measure includes questions and lexicalized responses (e.g., “Yes” and “I see”) as well as content-generating statements (i.e., topic extensions), as they demonstrate participation in the activity. This measure could also be used to triangulate the participants’ self-reported degree of extraversion. For example, if any participants rated themselves highly for extraversion but did not contribute much during the discussion, that would appear to present a contradictory finding.

Table 5. *Group-Based Fluency Measures*

Focus	Measure
Individual	Number of fluent syllables produced per minute
Group	Number of gaps between turns
	Mean length of gaps between turns

Another way of investigating fluency in a group-based context is to examine gaps between speaking turns. Efficiency of turn-taking contributes to the co-construction of fluency in interaction (Kasper, 2006), which then has to be collaboratively maintained (Peltonen, 2017). Inter-turn gaps are therefore not the responsibility of any individual and can only be calculated as a group statistic. Accordingly, the number and mean length of gaps between speaking turns were calculated in this study. In this case, however, only gaps of more than one second (silent or filled) were counted so that they were clearly noticeable and that other participants were likely to be aware that the previous speaker’s turn had ended. Moreover, if the same speaker followed a gap of more than three seconds with a different kind of interactional move it was counted as a new turn because it was felt that three seconds allowed enough time for the speaker’s previous turn to have clearly lapsed. Although this number is somewhat arbitrary, a standardized number was necessary to facilitate statistical comparisons and there were no previous studies to refer to in this regard.

Another important, albeit subjective, aspect of fluency is the effect that a sample of speech has on the listener. Perceived fluency constitutes the third element of Segalowitz's (2010) triadic model and cannot be overlooked given the inherently social, interactive nature of speech in most situations. Therefore, in addition to the objective measures of utterance fluency outlined above, the performance of each participant during the group discussions was assessed by raters on a five-point analytic scale (see Appendix R). The fluency descriptors focused on speech rate and hesitation phenomena given that these aspects of fluency have consistently shown a positive relationship with perceived fluency (Suzuki et al., 2021). The raters were language practitioners familiar with hearing Japanese learners of English, although they were not experts in the field of second language fluency so that any pre-conceived notions of the construct would likely be closer to those of the wider population. Training was provided in the form of one practice discussion, which the raters used to familiarize themselves with the rating scale and to standardize their scoring before assessing the speech samples used in the study.

As well as the participants' fluency, the raters assessed their communicative adequacy. It is possible to imagine speech that is superficially fluent, but difficult to understand or inappropriate in terms of efficiently achieving the goals of the task (Pallotti, 2009). Such speech could not be described as functionally adequate, so it is necessary to confirm that any gains in perceived fluency do not come at the expense of communicative adequacy. The descriptors were designed to reflect the main goals of the interactive task, which is to make relevant, comprehensible contributions with an appropriate level of detail.

Quantifying Interactional Competence

The recorded data from all the group discussions were also transcribed in Microsoft Word. All interactional micro-features used by the 20 core participants in each discussion were then coded based on the system used by Galaczi (2014). However, this coding system was adapted to include categories that have emerged from my 19 years' experience teaching EFL speaking and discussion classes in Japanese universities (see Tables 6 and 7). The coding scheme was intended to be exhaustive and account for all verbal features of interaction in this context. All coded features were then tallied in preparation for quantitative analysis. Non-verbal features, such as laughter or use of eye contact, were not quantified but were transcribed and addressed during qualitative analysis where appropriate.

Recognizing the listener and speaker roles in this academic EFL context, in which students are expected to give and support opinions, as well as respond to the opinions of others, I have operationalized interactional competence as consisting of *topic development* and *support moves*. The choice of these two coding domains follows Galaczi (2014), who was able to differentiate between speaking proficiency levels using a similar system. Topic development moves include both comments and questions that drive the topic forward, whereas support moves include backchannels and confirmations of comprehension that support the speaker and minimize misunderstandings, thereby maintaining mutual understanding.

Table 6. *Topic Development Moves*

Current study	Galaczi (2014)
Turn/topic management	Topic initiation
Topic extension (statement)	Topic extension (own)
Topic extension (question)	Topic extension (other)
Minimal extension	Minimal extension
	Echo

Table 7. *Support Moves*

Current study	Galaczi (2014)
Backchannels	Backchannels
Confirmations of comprehension	Confirmations of comprehension
Agree/disagree markers	
Repairs of misunderstandings	

However, given the highly contextualized nature of interactional competence, some features of interaction are likely to be task-dependent (Iwashita et al., 2021). In other words, features that are fundamental to one type of interaction might not occur in another context. For example, some backchannels (e.g., "Uh-huh" and "Yeah") are less relevant during single-unit than multi-unit turns, as their purpose is often to prompt the speaker to continue an incomplete turn (Schegloff, 1993). Therefore, any attempt to code and quantify interactional competence, whether for assessment or pedagogical purposes, is contingent on the specific context being investigated and should be tailored accordingly. For example, Galaczi's (2014) study was based on a high-stakes paired speaking exam (the Cambridge FCE exam), whereas in the current study I investigated unassessed group discussions in a Japanese, university-based EFL context. Although there are some underlying similarities between the two, such as the need to generate content and maintain comprehension among participants, they are fundamentally different types of interaction. For example, not only is the number of participants different (two versus three), but the lack of external assessment in these group discussions reduced the need for participants to display their competence, which could have resulted in less competition for speaking turns. Therefore, the contexts are sufficiently distinct to justify a modified coding system, as presented in Tables 6 and 7. See Appendix S for examples of the topic development and support features.

The type of language produced in any activity is constrained by the requirements of the task (Gan, 2010); therefore, the fact that the instructional demands of these two types of tasks also differ provides further justification for adapting the coding system. In the section of the test focused on by Galaczi (2014), candidates are typically provided with a variety of options to choose from (e.g., items to use on a desert island), along with picture prompts. These choices afford many opportunities to initiate a new topic, in addition to expanding on a previously initiated one, encouraging a transactional form of interaction. In contrast, the group discussion tasks in the current study consisted of two open-ended questions based on one pre-determined topic, such as high school life (Appendix J). Students were not provided with options but were required to formulate and support their own opinions related to each question. As suggested by the limited number of prompts, this kind of task calls for participants to discuss a topic in depth rather than continually initiate new ones.

Because the burden of topic initiation was removed in these discussion tasks, it is not included among the topic development moves to be coded (see Table 6). Nevertheless, the order in which students speak and whether consistent patterns are maintained in this regard is still of interest as the development of interactional competence has been described as a gradual increase in the ability to take the initiative (Pallotti, 2019). Volunteering to answer a question prompt first is one form of taking the initiative; therefore, this behavior was considered during the qualitative analysis.

As well as topic initiation, constraining the topics to be discussed also renders the categories of *own* and *other* topic extension moot because for a contribution to be relevant, it has to be related to that pre-determined topic (Grice, 1975). This issue is further complicated by the fact that topic shifts in conversation often lack clearly defined boundaries. An integral feature of spoken interaction is that participants link each turn to a previous turn so that topic transitions occur in a stepwise rather than disjunctive fashion (Holt & Drew, 2005). As a result, distinguishing between topics for analytic purposes becomes problematic because the boundaries between them are difficult to determine.

In the group discussion tasks in this study, participants can either develop the topic by making a statement related to it (e.g., "In my opinion, university life is more enjoyable than high school") or by asking fellow participants to do so. Questions could be directly related to a previous speaker's comment (e.g., "What kind of club did you join?") or related to the topic more generally ("Does anyone belong to a university club?"). Both kinds of questions demonstrate the ability to respond contingently—assuming they are related to what was said previously—which is a sign of higher speaking proficiency (Gan, 2010). Asking questions also entails taking responsibility for driving the discussion forward and can function as a way of extending a previous speaker's contribution. Patterns of interaction that include relevant and spontaneous questions, allowing the group to critically explore ideas, can therefore be seen as more developed than mechanical, pre-determined sequences of individual comments

(Nakatsuhara, 2011). Accordingly, a participant who encourages others to expand on their comments (e.g., "Why do you think so?") has displayed greater interactional competence than someone who engages only in mechanical interaction or waits passively for their turn to speak (Brouwer & Wagner, 2004). Making statements versus asking questions can, therefore, act as a proxy for extending one's own versus extending someone else's contribution. Consequently, I use the categories of *topic extension (statement)* and *topic extension (question)* rather than *topic extension (own)* and *topic extension (other)* for coding purposes. A topic extension statement can be spread over several turns, however, if listeners interject with confirmation requests or agreement markers, for example.

The two other topic development moves identified by Galaczi (2014) are *minimal extension* and *echo.* I have retained the former category, limiting it to statements of less than a clause (e.g., one-word answers to questions). However, I have not included response tokens such as "Yeah" and "Okay" unless they are direct answers to topic extension questions because they usually function as a form of backchanneling. For example, in response to the question "Have you joined any clubs?", "Yes" would count as a minimal extension, but in response to the statement "I joined the soccer club", "Yes" would count as a backchannel. In the latter case, it belongs in the domain of support rather than topic development moves because it provides no new content and therefore does not move the topic forward. Regarding echoing (i.e., when a listener repeats a word or phrase used by a speaker), if the word is repeated without the rising intonation used when questioning, its function is usually to acknowledge and encourage the speaker to continue (Riggenbach, 1991). Therefore, such tokens are counted as instances of backchanneling, unless they are specifically used to confirm comprehension in which case they are placed in that category (see Table 7).

Another difference between pair and group speaking activities is the regulation of turn-taking, which is inherently more complex in multi-party discourse. Interacting with a single interlocutor theoretically affords more opportunities for participation, whereas, logically, the presence of several interlocutors limits such opportunities, although this picture is complicated by personality and proficiency factors. As an essential element of group interaction, turn-taking management needs to be reflected in this coding system, so I have added the category of *turn/topic management* under topic development moves. Turn management refers to transitions between speaking turns, whereas topic management refers to transitions between the items to be discussed. These transitions are more relevant to group than pair interaction because the presence of more participants requires greater negotiation of roles. They maintain what Ducasse and Brown (2009) called horizontal and vertical cohesion; an example of the former (i.e., between speaking turns) is "Can I say something?" and of the latter (i.e., between topics) is "What shall we discuss next?"

Turn and topic management moves are distinct from the other types of topic development moves in that they do not directly extend the topic but rather facilitate transitions between speakers and topics. In other words, they function as precursors to topic extensions, thereby justifying their inclusion in this domain. Minimal replies to turn-taking questions such as "Sure, go ahead" are counted as backchannels, as they serve a similar purpose to other response tokens in that they invite the speaker to continue.

I have used turn and topic management differently from Galaczi (2014), who included a separate domain of *turn-taking management*, referring to the temporal coordination of turns in terms of gaps, latches, and overlaps. The timing of turn-taking is an important feature of interaction, but it is also related to the concept of fluency (in its broad sense) in terms of the overall smoothness of the discourse (Riggenbach, 1991). Nevertheless, so as not to conflate the measurements of interactional competence and fluency, it is necessary to place this aspect of turn-taking in either one domain or the other. No other elements of interactional competence were measured temporally in this study; therefore, the timing of turn-taking was quantified under group-based fluency, as a measure of the length of gaps between speaking turns (see Table 5). However, as an important element of interactional competence, the temporal aspect of turn-taking was addressed during qualitative analysis.

As well as providing or requesting content to develop the topic, the other main role of participants in a group discussion, given its social nature, is to support the other members (Dörnyei & Murphey, 2003). However, speaker and listener roles are in a constant state of flux during social interaction (Sert, 2019), so I have given this domain the generic label *support moves*, rather than *listener support moves* (in contrast to Galaczi, 2014). This domain includes the category of *backchanneling*, which includes response tokens such as "Yes" and "Okay" (unless they are direct answers to topic extension questions), and echoes (unless they are used to confirm comprehension). Such reactions perform an important role in indicating support for other participants; furthermore, their use is perceived favorably by highly proficient listeners when rating learner interaction (Brown et al., 2023).

For quantification purposes, I counted multiple backchannels of the same type within the same topic extension statement as only one instance of backchanneling, in order not to distort the statistics. For example, “Yes, yes, yes” and “Okay” would only count as one backchannel, even if they did not appear in the same turn because they are both simple response tokens. However, a subsequent echo within the same topic extension statement would be additionally counted. Non-lexical backchannels, such as “mm-hmm” and “uh-huh”, were not counted, as they require little alignment, or even comprehension, and are often difficult to attribute to individuals during group-based interaction. This issue is heightened when participants are wearing face masks, as many were during the current study due to the lingering effects of the COVID-19 pandemic.

The second category of support moves is *confirmations of comprehension*. This category clearly demonstrates the intertwined nature of speakership and listenership, as confirmations of comprehension can come either from the speaker’s perspective (e.g., “Do you understand?”) or listener’s perspective (e.g., “Yes, I understand”). However, they have to be intended for the purpose of confirming comprehension to be counted in this category. Unsolicited instances of “I see” are counted as backchanneling if there is no prior indication of a threat to comprehension.

Repairs of misunderstandings (e.g., “I mean...”) also come under the domain of support features, whether they involve initiating or responding to other-initiated repair. However, they are only counted in this category if it is clear that a breakdown in understanding has taken place. For example, a listener checking that they have understood an utterance by asking “Do you mean...?” counts as confirming comprehension rather than repairing misunderstanding. Nevertheless, a repair could then ensue if it becomes apparent that the person confirming has indeed misunderstood the speaker's original comment.

Indicating agreement or disagreement with a previous speaker's comment also comes under the domain of support features. Galaczi (2014) included tokens of agreement (e.g., "Exactly") within the category of confirming comprehension, although such tokens perform a different role because they demonstrate not just comprehension but also alignment and mutuality with the previous speaker. They are also distinct from backchanneling tokens in that they display greater involvement and engagement with a previous utterance (Wong & Waring, 2021). For these reasons, and the fact that they are fundamental to opinion-based discussions, I have placed indicators of agreement or disagreement in a separate category, which I have called *agree/disagree markers*. However, to count as such, the utterance must include an element of agreement or disagreement (e.g., *I think so too* or *I see your point, but...*) and not just a reaction such as surprise (e.g., *Wow!*), which would count as a backchannel. An agree/disagree marker can be followed by a topic extension comment if further content is added that moves the discussion forward.

To summarize, the interactional features to be quantified in this study are divided into two domains: topic development moves and support moves. The former domain consists of turn and topic management, topic extension (statement), topic extension (question), and minimal extension. The latter domain consists of backchannels, confirmations of comprehension, agree/disagree markers, and repairs of misunderstandings.

Qualitative Analysis

As previously mentioned, quantification highlights patterns of interaction among multiple participants over time, although this analysis must be complemented by a qualitative examination of the data. In other words, the quality of responses is also of critical importance, as a quantitative increase in the use of a feature does not necessarily imply improvement (Lam, 2018). Moreover, qualitative analysis reveals whether participants have diversified their interactional techniques within the same quantified category. For example, if two participants both use the same number of agreement markers but Participant A uses the same form each time while Participant B uses several different forms, Participant B has displayed a greater diversification of methods, which is a key component of interactional competence development (e.g., Pekarek Doehler, 2019). Furthermore, qualitative analysis can also show the context in which each form was used in order to shed light on its appropriacy in that context and how it was oriented to by other participants.

For the qualitative analysis in this study, exemplars are presented of the interactional features that have been quantified using the coding system outlined above. These exemplars were taken from the initial transcriptions of the group discussions and re-transcribed in more fine-grained detail using conventions informed by CA (Atkinson & Heritage, 1984). A similar system was used by Galaczi (2014), who used these conventions to examine micro-level interactional features in the context of paired speaking tests. Galaczi (2014) presented the speech in two columns, "to illustrate more clearly individual contributions to the joint discourse" (p. 560). However, as I am presenting data from three-member discussions, the use of columns for individual speakers would not work in terms of page layout. Therefore, I used a more conventional line-by-line approach to transcription. The choice of transcription conventions used in this study is shown in Appendix T.

Instances were sought of participants using interactional strategies, especially those taught during the course (e.g., paraphrasing), illustrating the reflexive relationship between pedagogy and interaction (Seedhouse, 2004). How, if at all, these strategies facilitated the development of interactional competence could then be investigated. For example, “What do you think?” is a more precise way of eliciting an opinion than “How about you?”, which can be used to elicit either a fact or an opinion and requires the recipient to decide which was intended. Both questions would be coded as instances of turn-taking management, although the former reduces any threat to comprehension, due to its greater specificity. In other words, it demonstrates greater sensitivity to issues of context and recipient design, which Pekarek Doehler (2019) has argued is at the heart of L2 interactional competence development. Therefore, a participant who progresses from using the general form to the more precise form has arguably improved their interactional competence in the context of a group discussion. Examining such exemplars through a system informed by CA illustrates how interactional competence develops by comparing, in fine-grained detail, output from the beginning, middle, and end of the course. As succinctly stated by Wong and Waring (2021), “Conversation analysis delivers the stuff that interactional competence is made of, i.e., interactional practices” (p. 8).

Although quantitative and qualitative analysis emanate from different research traditions, I believe they offer complementary, rather than contradictory, perspectives. In what Seedhouse (2005) has referred to as a “linguistic CA approach” (p. 175), qualitative analysis can be used to isolate interactional features, which can then be treated quantitatively in extensive data sets. If this approach is adapted to the local context, as in the current study, it can help to resolve the fundamental tension at the heart of interactional competence research; that is, the need to define a stable construct that can be taught with the need to preserve the contingency and variation inherent in specific contexts (Lee, 2006). Therefore, combining the two research paradigms, especially in an exploratory study such as this, can help to illuminate the development of interactional competence.

Analyses

To answer the first research question, which asked the extent to which fluency developed during a one-semester course of spoken interaction, repeated measures ANOVAs were used to examine changes in monologue utterance fluency among the 20 core participants. Regarding the group discussion-based fluency measures (see Table 5), a repeated measures ANOVA was used to examine changes in the number of fluent syllables produced per minute. Descriptive statistics alone were used to analyze mean length and number of gaps between turns, as each group consisted of different members at each time point, which precluded the use of a repeated-measures ANOVA.

Regarding the perceived fluency ratings, the validity of all raters' scores was investigated using Rasch FACETS analyses (see Appendix U for an example FACETS specification file). The many-facets Rasch model is able to control for factors that have a systematic influence on the data. This characteristic is important because raters vary in their severity when using a rating scale. Fit statistics were therefore computed to examine how consistently each rater assessed the participants' fluency and communicative adequacy. Using logits produced by the Rasch analysis, repeated-measures ANOVAs were then conducted to ascertain the extent of any changes over time. Correlation analysis was also conducted to investigate whether these subjective measures were related to the objective measures of utterance fluency and/or number of fluent syllables produced in the group discussions.

To answer the second research question, which asked the extent to which interactional competence developed during a one-semester course of spoken interaction, both quantitative and qualitative methods were used. Changes in the use of the interactional features (topic development and support moves) over the three measurement points were examined. However, the number of moves produced by an individual in one discussion was likely to be small; therefore, it was unlikely that the data would be normally distributed. Furthermore, the fact that these moves are contingent on the behavior of other participants means that each measure was not independent, so the use of descriptive and non-parametric statistics—in this case Friedman's repeated measures ANOVA—was appropriate (Turner, 2014). Qualitative analysis was then used to examine the use of these features in detail. The highly proficient speaker discussion was also analyzed in terms of the same aspects of interactional competence. These data were then compared with that produced by the learner discussions at each measurement point to ascertain whether the interactional features of learner discussions had progressed towards the performance of highly proficient speakers.

Questionnaire data were used to answer the third research question, which asked the extent to which speaking self-efficacy developed during a one-semester course of spoken interaction. The participants were asked to rate their feelings of self-efficacy and anxiety when using English in a range of situations related to this learning context, at the beginning, middle, and end of the course. The results of these questionnaires were analyzed via a repeated measures ANOVA, using logits produced by the Rasch analysis, allowing the magnitude of changes in speaking self-efficacy and anxiety during the course to be observed.

To answer the fourth research question, which asked the extent to which fluency was related to interactional competence, correlation analysis was used. All aspects of an individual's fluency included in this study—as measured in monologue activities, group discussions, and as perceived by native-speaker raters—were correlated with the raw scores of the interactional features at all three measurement points. However, to avoid the risk of a Type I error (i.e., to erroneously conclude the significance of a positive correlation) a single measure of monologue utterance fluency—pruned speech rate—was used. In addition, the eight categories of interactional features were aggregated into two domains (i.e., topic development and support moves) for the purposes of this analysis. Table 8 shows the variables included in this correlation analysis. All measures pertain to the individual; however, all except pruned speech rate were taken from the group discussion activities.

Table 8. *Variables Included in Correlation Analysis (Research Question 4)*

Category	Measure
Fluency	Pruned speech rate (monologue)
	Number of fluent syllables per minute (group)
	Perceived fluency (group)
Interactional competence	Topic development moves
	Support moves

Correlation analysis was also conducted to answer the fifth and sixth research questions, which asked the extent to which fluency and interactional competence were related to the individual difference variables of speaking self-efficacy, anxiety, and extraversion—also measured using questionnaire data. For research question 5, once again all aspects of an individual's fluency were included, with pruned speech rate the only utterance fluency measure used. For research question 6, the aggregated scores of topic development and support moves were used. Tables 9 and 10 show the variables included in these correlation analyses. The analysis was conducted at all three measurement points, although extraversion was only included at the first measurement point as that questionnaire was only distributed once.

Table 9. *Variables Included in Correlation Analysis (Research Question 5)*

Category	Measure
Fluency	Pruned speech rate (monologue)
	Number of fluent syllables per minute (group)
	Perceived fluency (group)
Individual difference	Speaking self-efficacy
	Speaking anxiety
	Extraversion

Table 10. *Variables Included in Correlation Analysis (Research Question 6)*

Category	Measure
Interactional competence	Topic development moves
	Support moves
Individual difference	Speaking self-efficacy
	Speaking anxiety
	Extraversion

In addition, for the speaking self-efficacy, anxiety, and extraversion questionnaire items, the Rasch measurement model (Rasch, 1960), using WINSTEPS version 5.6.0 (Linacre, 2023), was used to ensure that all items demonstrated sufficient fit, unidimensionality, and reliability to be included in the study (see Appendix V for an example WINSTEPS control file).

CHAPTER 4

PRELIMINARY ANALYSES

In this chapter, I introduce the Rasch measurement model (Rasch, 1960), then describe how it was used to examine reliability and validity of the individual difference questionnaires and rating scales. Regarding the individual difference variables, the dimensionality and model fit of the extraversion, speaking self-efficacy, and speaking anxiety measures were investigated using WINSTEPS version 5.7.1 (Linacre, 2023). In addition to model fit, the reliability estimates of the perceived fluency and communicative adequacy ratings were investigated using FACETS version 4.1.4 (Linacre, 2024).

The Rasch Model

Rasch model analyses were conducted to investigate the validity of several measures in this study: the extraversion, speaking self-efficacy, and speaking anxiety questionnaires, and the scoring rubric used by the raters to measure perceived fluency. There are several benefits to using the Rasch model to examine the validity of these measures. First, the raw scores or ratings can be converted from ordinal to interval scales, which allows statistical calculations to be performed on the data. For example, it cannot be assumed that the difference between *Strongly agree* and *Agree* is the same as that between *Agree* and *Moderately agree* simply because they are adjacent categories on the Likert scale. In other words, it might require proportionately more of the construct under investigation to move between thresholds at different points in the scale. Rasch analysis uses statistical probability to create logits—short for log odds— based on interval measures, which can then be used for the repeated measures ANOVA and correlational analysis mentioned in the previous section.

Another advantage of Rasch analysis is that fit indices can be produced to determine the degree to which the person and item data fit the model. Specifically, if the amount of variation between the observed data and the responses predicted by the model are widely divergent, the validity of the measures is threatened. For example, misfitting questionnaire items undermine the principle of unidimensionality, as such items might not be related to the underlying construct under investigation (e.g., extraversion). Regarding person data, inconsistent patterns of behavior (i.e., when answering questions or rating participants) could indicate a lack of understanding or attention. On the other hand, insufficient variation among an individual's responses could imply a holistic, rather than an analytic, approach has been applied (Bond et al., 2021). The mathematical principles underlying Rasch analysis are outlined below.

Dichotomous versus Polytomous Data

The Rasch model was originally conceived of in relation to dichotomous data, such as a test question that has either a correct or an incorrect answer. In this case, the model estimates the probability of a test subject (i.e., a person) correctly answering a test item in the following way. A person estimate is obtained by calculating the ratio of items previously answered correctly by that individual, whereas an item difficulty estimate is obtained by calculating the ratio of other subjects who correctly answered the same item. Then, using these two values, the probability of a person answering an item correctly can be calculated using the following equation:

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

where P_{ni1} is the probability of person *n* answering item *i* with a score *(x)* of 1, which is a function *(f)* of the person's ability (B_n) minus the difficulty of the item (D_i).

However, the Rasch model can also be extended to polytomous data, which involve more complex responses than a simple right or wrong distinction. When applied to Likert scale data, as in the current study, the rating scale model (Andrich, 1978) creates difficulty estimates for each threshold between the item response categories (e.g., *Strongly agree* and *Agree*), as well as a set of shared rating scale thresholds that is common to all items. Similar to item measures in the dichotomous model, polytomous item thresholds are modeled as the probability of a participant choosing one response category over another, which can be stated as:

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

where P_{ni1} is the probability of person *n* selecting the response category 1 instead of the first response category 0 on item *i*. Here, D_i represents how difficult the item is to endorse, and F_1 represents the difficulty of endorsing it at the level of the first threshold in the response category.

Validating a rating scale used for scoring purposes, such as the perceived fluency rubric in this study, adds even more elements to the model, further increasing its complexity. For example, raters often vary in terms of their severity, which introduces additional noise that needs to be accounted for. Many-facet Rasch measurement (MFRM) controls for such factors, as long as they influence the data systematically, as it allows for the simultaneous analysis of multiple variables—or facets—to be calibrated on a single interval scale. When calculating the probability estimate of person *n* responding to item *i* for a category threshold k, the severity of rater j is taken into account. This formula can be expressed as:

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

where these estimates are a function of the ability of person B_n, the difficulty of item D_i, the difficulty of the threshold F_k, and the severity of the rater C_j. The facets under investigation in the current study are: person ability (i.e., based on the scores assigned by the raters), rater severity, and task difficulty (i.e., the three different discussion topics). In common with other Rasch models, MFRM also provides a set of indices that can be used to determine the extent to which the data fit the Rasch model.

Fit Indices

Given that Rasch analysis is based on statistical probability, it is unrealistic to expect the observed data to align perfectly with the estimates produced by the model. Nevertheless, for a scale to be validated it is essential to identify any persons and items that behave in unpredictable or otherwise undesirable ways. This information can then be used to determine whether such data are effectively contributing to measurement of the underlying construct of interest. In other words, fit indices can be used to judge how well the data fit the model, thus operating as a form of quality control (Bond et al., 2021). In short, excessively predictable responses *overfit* the model, whereas erratic and inconsistent responses *underfit* the model.

Mean square outfit statistics are unweighted averages of the residual variance (Z_{ni}) across persons and items and are useful for identifying outlying data points. In mathematical terms, they can be stated as:

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

Mean square infit statistics, on the other hand, are residuals weighted in favor of unexpected responses that occur near a person's ability estimate or an item's difficulty estimate. As such, infit values provide a more sensitive reflection of an item or person's performance and are therefore often prioritized over outfit values when assessing item fit (Bond et al., 2021). As weighted averages of the residual variance (W_{ni}) across persons and items, infit statistics can be stated as:

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

Infit and outfit mean-square statistics have an expected value of 1, although the intended purpose of the results must be considered when determining the range of acceptable values. For example, high-stakes tests call for tight quality control limits on the items used—in this case Bond et al., (2021) recommend 0.80 to 1.20—although in lower stakes or exploratory contexts more generous limits of 0.50 to 1.50 have been shown to be productive for measurement purposes (Linacre, 2002b). Therefore, as a classroom-based exploratory study, the latter range was deemed appropriate for the current context. Fit statistics can also be provided on a standardized scale (*Z*STD). Values above 2.00 are considered too erratic to be of use, whereas values below -2.00 demonstrate insufficient variation (Bond et al., 2021).

Reliability and Separation Indices

Another advantage of Rasch analysis is that it indicates how reliably persons and items are separated by a scale into distinct levels of performance or difficulty. Person separation, for example, indicates the extent to which participants differentially exhibit the construct of interest. The spread of person or item data is provided by separation indices, with person separation (G_p) calculated as follows:

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

where the adjusted person standard deviation (SA_p) is divided by the average measurement error (SE_p). Items can similarly be separated in terms of difficulty by substituting item for person variance.

Reliability, on the other hand, indicates how replicable the results are, with an estimate's values ranging between 0.00 and 1.00. Values closer to 1.00 indicate that results are likely to be replicable if a parallel version of the test or questionnaire were to be conducted with the same participants, or if different participants were presented with the same items. This estimate can be calculated as follows:

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

where the variance of the measures (SA_p^2) is divided by the total person variability (SD_p^2).

Rating Scale Functioning

An additional purpose of Rasch analysis is to determine whether rating scales are functioning appropriately. To this end, Linacre (2002a) provided a set of criteria that can be used to evaluate the effectiveness of a rating scale. First, there should be a minimum of ten observations per category, with observations distributed regularly across categories. Next, average category measures should increase monotonically in accordance with the rating scale categories, and respondents who exhibit more of the construct should be observed in higher categories of the rating scale. Also, outfit mean-square statistics should be under 2.00, as values above this number indicate the presence of more unexplained than explained variance. Finally, Rasch-Andrich category thresholds should be between 1.40 to 5.00 logits for scales consisting of three category levels. It has since been calculated that the minimum separation for larger scales—such as the questionnaires in this study—should be reduced to approximately 1.10 and 0.81 logits for four and five categories, respectively (Wolfe & Smith, 2007), and 0.59 logits for scales with six categories (Elwood, 2011).

The distinctions between rating scale category thresholds can also be visually inspected via probability curve graphs, which illustrate the probability of endorsing any given category for every difference estimate. Each category should have a distinct peak in the probability curve graph, demonstrating that it is the most probable response category for some portion of the measured variable (Bond et al., 2021). Any category that does not span an observably distinct portion of the variable is therefore redundant, so it might be necessary to collapse adjacent categories together so that the model conforms to expectations.

Unidimensionality

Rasch analysis is based on the concept of unidimensionality, which requires that the items or rating criteria entered into the model all tap into the same latent variable. The dimensionality of items can be investigated via principal components analysis (PCA), whereby standardized residuals are checked for the existence of non-random variance. Factor loadings of above .40—whether positive or negative—indicate the possible existence of additional dimensions (Wolfe & McVay, 2012). However, as Eckes (2015) has pointed out, it is not necessary for all items to be based on a single psychological component; rather, the items should work together to form a single pattern of observations. In other words, the items should form a psychometrically, rather than psychologically, unidimensional construct. Nevertheless, if significant amounts of such variance can be attributed to factors outside of the variable of interest—for instance gender, ethnicity, or L1—multidimensionality is implied and construct validity undermined. Many researchers (e.g., Bond et al., 2021) are reluctant to state unambiguous cut-off points for unidimensionality, emphasizing instead the need to examine items that load substantially on secondary dimensions and to consider their relationship with the construct. However, according to Linacre (2017), a construct can probably be considered sufficiently unidimensional for Rasch analysis purposes as long as no secondary dimension has an eigenvalue above 3.00. Also, the percent of variance accounted for by the first residual contrast should ideally be less than 10%.

An additional indication of dimensionality is found in item polarity, which is revealed by the point-measure correlation. This statistic provides information regarding the relationship between items and is calculated by correlating individual item measures with the combined measures of the other items. High values for these correlations indicate that the measures are well aligned and are working in the same direction, whereas negative or close to zero values imply significant deviation from the model's expectations. In general, values above .40 are considered to indicate consistency of item polarity (Bond et al., 2021).

Rasch Analysis of the Extraversion Questionnaire

The extraversion questionnaire (Appendix O) was administered at the first data collection point only, as personality traits are considered relatively stable and therefore unlikely to change during one semester. The eight items were entered into a Rasch rating scale analysis; the 20 participants thus created a data set of 160 individual item responses. Initially, a six-point Likert scale was used, whereby 1= *Strongly disagree*, 2 = *Disagree*, 3 = *Moderately disagree*, 4 = *Moderately agree*, 5 = *Agree*, and 6 = *Strongly agree*. However, not all the Likert-scale categories were found to be functioning in accordance with the criteria outlined above. Specifically, there were only four counts of *Strongly disagree* in total, and the separation between Andrich thresholds was only 0.50 between categories 2 and 3. Therefore, the two lowest categories—*Strongly disagree* and *Disagree*— were collapsed into one *Disagree* category. The results of this reduced five-point scale are shown in Table 11. The minimum of 10 observations per category was now met, with precisely that number observed in Category 5. The average measure increased in accordance with the rating scale categories and the outfit mean square (MNSQ) statistic was well below 2.00 for all categories. The Andrich thresholds advanced as expected, with separation between categories now comfortably within acceptable limits. The corresponding category probability curves are shown in Figure 4, which demonstrates how the rating scale accords with the expectations of the Rasch model. After examining these data, it was concluded that the adapted Likert-scale was performing in a satisfactory manner.

Table 11. *Category Structure Functioning for Extraversion*

Category	Observed count (%)	Average measure	Outfit MNSQ	Andrich threshold	Threshold change
1. Disagree	15 (9)	-3.46	0.95	—	—
2. Moderately disagree	54 (34)	-1.73	1.06	-3.94	—
3. Moderately agree	50 (31)	-0.19	1.04	-0.84	3.10
4. Agree	31 (19)	1.81	0.78	1.34	2.18
5. Strongly agree	10 (6)	2.96	1.11	3.45	2.11

Note. All statistics are based on Rasch logits.

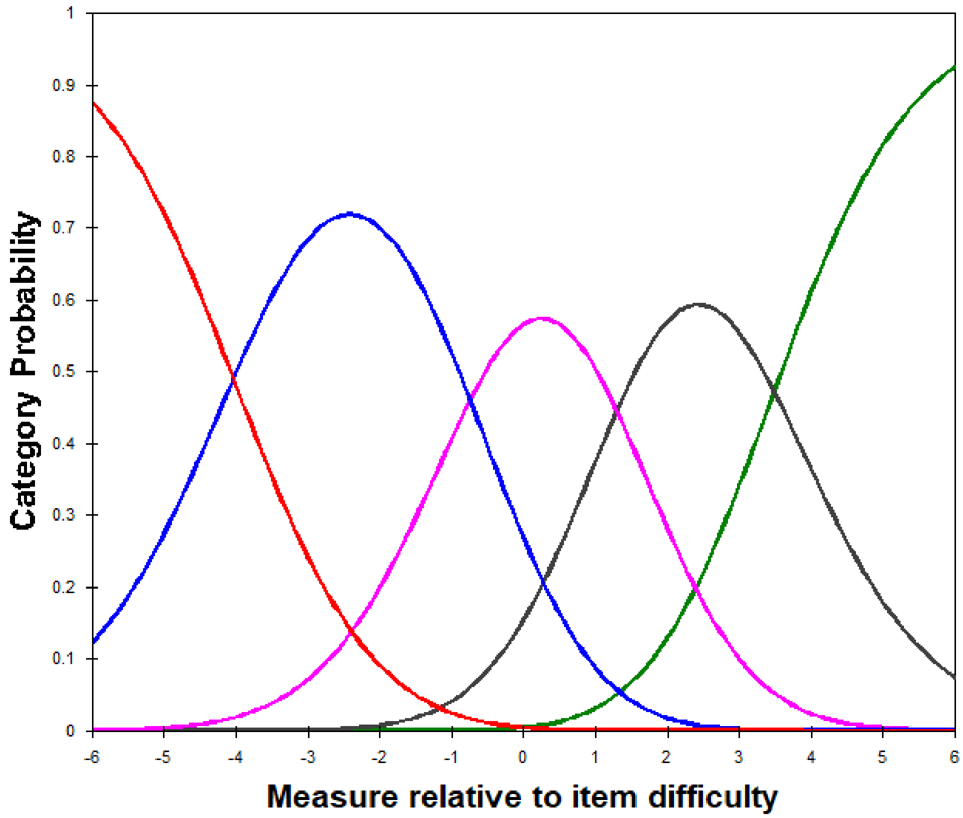

Figure 4. Rating Scale Performance of Extraversion Measured on a 5-Point Likert Scale

Item Fit and Endorsability

Table 12 shows the fit statistics for the questionnaire items, revealing that all eight items were within the acceptable outfit MNSQ range of 0.50 to 1.50 (Linacre, 2002b). However, item EXT7, *I feel at ease in a large group of people*, has an infit MNSQ value of 1.51, which is just beyond the acceptable limit, although its ZSTD infit and outfit values are both below 2.00, indicating compatibility with the model. Furthermore, its point-measure correlation was above the suggested threshold of .40 (Bond et al., 2021). The item was therefore retained in the subsequent analyses.

Table 12. *Rasch Item Statistics for the Extraversion Items*

Item	Measure	Model *SE*	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt-measure correlation
EXT5	1.57	0.38	0.91	-0.20	0.85	-0.40	.71
EXT1	1.28	0.37	1.14	0.50	1.13	0.50	.79
EXT7	0.61	0.36	1.51	1.50	1.48	1.50	.56
EXT6	-0.17	0.36	0.77	-0.70	0.78	-0.70	.89
EXT8	-0.17	0.36	0.80	-0.60	0.83	-0.50	.78
EXT4	-0.29	0.36	1.11	0.50	1.10	-0.40	.77
EXT3	-0.92	0.35	0.78	-0.60	0.83	-0.50	.87
EXT2	-1.92	0.36	0.84	-0.40	0.89	-0.30	.78

Note. All statistics are based on Rasch logits.

The Wright map (Figure 5) shows the participants—displayed as Xs—on the left-hand side of the logit scale. Participants with higher levels of extraversion are found toward the top of the map and persons with lower levels of extraversion toward the bottom. The questionnaire items are displayed on the right-hand side of the figure according to their difficulty estimates, with more difficult to endorse items at the top of the map and easier to endorse items at the bottom. The average person measure was -.43, indicating that the items were reasonably well matched to the participants, although the Wright map shows that the participants are fairly evenly distributed throughout the scale. The extraversion measure had a reliability estimate of .87 for items and .88 for persons, while item separation statistic was 2.64 and person separation was 2.72. According to Fisher's (2007) scale, these results indicate good reliability and a fair degree of separation, although the latter is limited by the relatively small number of items and participants.

```
--------------------------------------------------------------------------------
More extraversion (persons) More difficult to endorse (items)
    3                 +
                  XX   |
                       |
                       |
                       |T
    2                  +
                       |
                   X   |  EXT5: Like to be center of attention
                  XX  S|
                       |  EXT1: Am the life of the party
    1              X   +S
                   X   |
                       |  EXT7: Feel at ease in large group of people
                  XX   |
                       |
    0                  +M
                       |  EXT6: Talk a lot  EXT8: Skilled at dealing with people
                      M|  EXT4: Get along with almost anyone
                       |
                 XXX   |
   -1                  +S EXT3: Start conversations
                       |
                  XX   |
                       |
                   X   |
   -2                  +  EXT2: Like to talk to a lot of people
                       |T
                   X  S|
                       |
                  XX   |
   -3                  +
                  XX   |
                       |

Less extraversion (persons) Easier to endorse (items)
--------------------------------------------------------------------------------
```

Figure 5. Wright Map Showing Extraversion Item-Person Responses
Note. M = mean, S = one standard deviation, T = two standard deviations, X = one student, *N* = 20. Measure values are in Rasch logits.

Regarding the hierarchy of difficulty, the items could be divided into roughly three levels based on their relative ease of endorsement. The most difficult items to endorse were EXT5, *I like to be the center of attention*, and EXT1, *I am the life of the party*. These two items represent a high degree of extraversion, as they suggest enjoying being the focus of attention among a group of people. In contrast, the easiest item to endorse was EXT2, *I like to talk to a lot of people*, which implies an enjoyment of socializing although not necessarily of drawing attention to oneself. Finally, the middle cluster of items, such as EXT6, *I talk a lot*, and EXT8, *I am skilled at dealing with people*, suggest that most participants felt moderately comfortable about interacting and dealing with people in general.

Dimensionality

In terms of dimensionality, the Rasch PCA of item residuals showed that 66.40% of the total variance was explained by the model, with an eigenvalue of 15.79. An additional 10.40% of the variance was explained by the first residual contrast, with an eigenvalue of 2.48. The value of this secondary dimension is above the level that can be attributed to random noise, although it is less than the practical criterion of 3 mentioned by Linacre (2017). Therefore, the construct was sufficiently unidimensional for Rasch analysis purposes. Table 13 shows the loading of item residuals from the extraversion questionnaire. Items EXT2 and EXT7 had high positive loadings, while items EXT1 and EXT5 had negative loadings below -0.40. Therefore, these items need to be examined to ensure that they are tapping into different aspects of the construct of extraversion rather than unrelated dimensions. In the positive case, the two items specifically mention large groups or a lot of people, so they could be addressing a subcomponent of interacting with many people simultaneously. In the negative case, the two items mention being the center of attention or the life of the party, suggesting a subcomponent of drawing attention to oneself. The remaining items, EXT3, EXT4, EXT6, and EXT8 were concerned with interaction in a more general sense, such as starting conversations and getting along with people. Therefore, the three subcomponents mentioned above could be considered somewhat distinct aspects of the underlying construct of extraversion.

Table 13. *Rasch PCA of Item Residuals for the Extraversion Items*

Item	Loading	Measure	Infit MNSQ	Outfit MNSQ
EXT2	.89	-1.92	0.84	0.89
EXT7	.71	-0.61	1.51	1.48
EXT3	.27	-0.85	0.83	0.89
EXT4	-.17	-0.29	1.11	1.10
EXT6	-.19	-0.17	0.77	0.78
EXT8	-.40	-0.17	0.80	0.83
EXT5	-.47	1.57	0.91	0.85
EXT1	-.81	1.28	1.14	1.13

Note. All statistics are based on Rasch logits.

Person Measure Descriptive Statistics

Following Rasch analysis, descriptive statistics for the extraversion person measures were examined. The results, using Rasch logits, can be seen in Table 14. Levels of skewness and kurtosis were within acceptable limits of -2.00 to +2.00 (Hair et al., 2022), although the latter value of -1.12 indicates that the data were trending towards a platykurtic distribution. However, the fact that the participant measures are spread relatively evenly suggests that there is a balance of high, mid, and low levels of extraversion among these 20 participants, which is useful for correlational analysis purposes. Person measures were also converted to *z*-scores, all of which fell comfortably within the range of -3.29 to 3.29, further confirming that the results for the extraversion measure were normally distributed with no outliers.

Table 14. *Descriptive Statistics for Extraversion Person Measures*

	Extraversion
M measure	-0.43
M SE	0.44
95% CI	[-1.28, 0.42]
SD	1.95
Skewness	0.07
SES	0.51
Kurtosis	-1.12
SEK	0.99

Note. *N* = 20. All statistics are based on Rasch logits.

Rasch Analysis of the Speaking Self-Efficacy Questionnaire

A similar procedure was followed to confirm the validity of the speaking self-efficacy and anxiety questionnaires. However, they were administered three times, rather than just once as in the case of the extraversion questionnaire, as affective individual differences are considered less stable than personality traits and can thus potentially change even over a relatively short period of time. The measures taken from all three administrations were entered into a single analysis using stacked participant codes, resulting in three rows of data for each participant. This analysis was then examined for model fit and category functioning. The self-efficacy questionnaire (Appendix K) consisted of eight items, although preliminary screening revealed that item SE5, *I can listen actively to my classmates during English speaking class*, misfit the model with infit and outfit MNSQ values above 2.00, and standardized ZSTD values above 4.00—more than double acceptable levels in the latter case. Upon closer examination, the fact that this item contains the word *listen*, albeit in the context of active participation, could have caused respondents to have interpreted it differently, leading to the erratic patterns outlined above. For example, some might have interpreted it as relating purely to listening ability, rather than the ability to respond actively while in the role of listener. The item was therefore removed from the analysis. The resulting data set consisted of 419 individual item responses, as one data point was missing from one of the questionnaires. The Rasch model is robust enough to not require complete data sets, however, so the absence of only one data point had a minimal impact on the estimated measures.

Again, a six-point Likert scale was initially used, whereby 1= *Strongly disagree*, 2 = *Disagree*, 3 = *Moderately disagree*, 4 = *Moderately agree*, 5 = *Agree*, and 6 = *Strongly agree*. However, as with the extraversion questionnaire, not all the Likert-scale categories were found to be functioning satisfactorily, with only six counts of *Strongly disagree* in total, and the separation between Andrich thresholds only 0.51 between Categories 2 and 3. Furthermore, after the removal of item 5, only seven counts remained of *Strongly agree*, and the separation between Andrich thresholds was above 5.00 between Categories 5 and 6. Therefore, the two lowest categories—*Strongly disagree* and *Disagree*—and the two highest categories—*Strongly agree* and *Agree*— were collapsed into one *Disagree* and one *Agree* category, respectively. The results of this reduced four-point scale are shown in Table 15. The minimum of 10 observations per category was now met, with the lowest being double that number observed in Category 1. Although there were a far higher number of responses in Categories 3 and 4, the average measure increased in accordance with the rating scale categories, the Andrich thresholds advanced, and separation between categories was now within acceptable limits. The outfit mean square statistic was also below 2.00 for all categories. The corresponding category probability curves are shown in Figure 6, demonstrating how the rating scale conforms to the model expectations. After examining these data, it was concluded that this adapted Likert-scale was performing in a satisfactory manner.

Table 15. *Category Structure Functioning for Speaking Self-Efficacy*

Category	Observed count (%)	Average measure	Outfit MNSQ	Andrich threshold	Threshold change
1. Disagree	20 (5)	-2.51	0.71	—	—
2. Moderately disagree	57 (14)	-0.42	1.02	-2.46	—
3. Moderately agree	187 (45)	1.65	0.81	-0.67	1.79
4. Agree	155 (37)	4.15	1.09	3.13	3.80

Note. All statistics are based on Rasch logits.

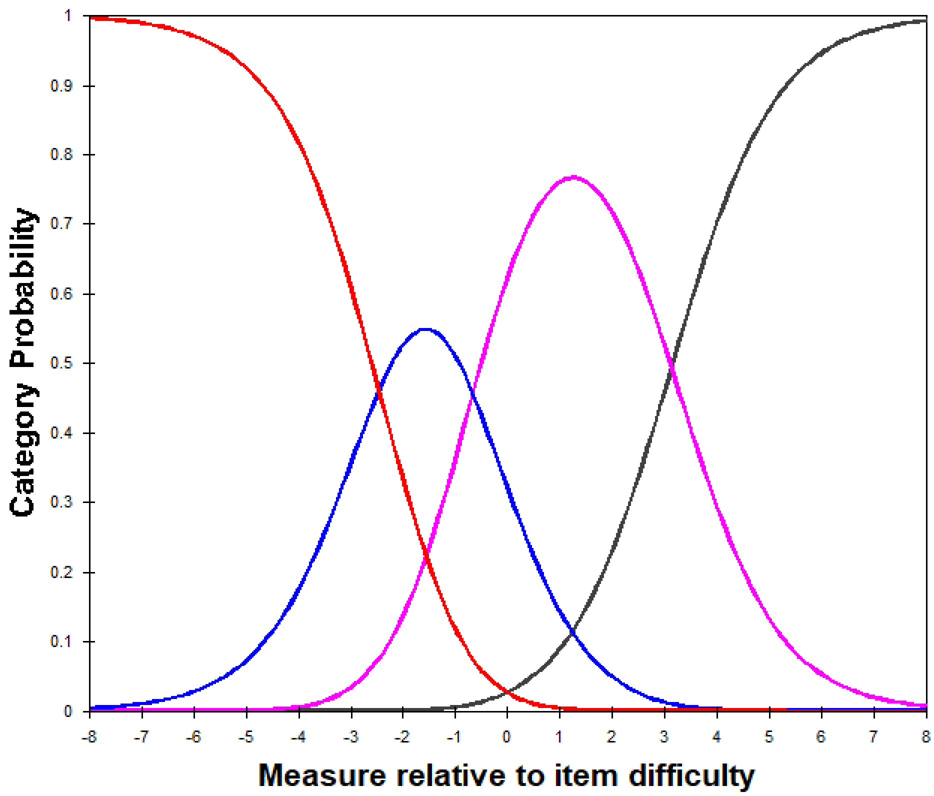

Figure 6. Rating Scale Performance of Self-Efficacy Measured on a 4-Point Likert Scale

Item Fit and Endorsability

Table 16 shows the fit statistics for the remaining seven questionnaire items, revealing that all items are within the acceptable outfit MNSQ range of 0.50 to 1.50. The average person measure was 2.14, indicating that the participants had quite a high degree of self-efficacy regarding the activities in this speaking class, which is confirmed visually by the Wright map (Figure 7). However, it is important to note that this analysis combines all three iterations of the questionnaire, so it is possible that this finding reflects changes in the participants' self-efficacy over time. The Rasch reliability estimate was .96 for items and .82 for persons, which according to Fisher's (2007) scale represents excellent and good, respectively. The item separation estimate was 4.64 and person separation was 2.12, representing very good and fair, according to the same scale.

Regarding the hierarchy of difficulty, *I can solve communication problems in English* (SE6), proved to be the most difficult item to endorse, followed by SE8, *I can speak English only for a whole lesson*. These are the two items that I had hypothesized would be the most difficult to endorse on the basis that the students would not be used to speaking only in English for such a long period of time, and that the need to solve problems implies that something had previously gone wrong during the interaction. The two easiest items to endorse were SE3, *I can speak English when talking to a partner in pair work activities*, and SE2, *I can use the phrases and skills that I learn in an English speaking class*. I had also hypothesized that SE3 would be relatively easy to endorse, on the basis that talking to only one partner is less face-threatening than talking in a group.

```
--------------------------------------------------------------------------------
More self-efficacy (persons) More difficult to endorse (items)

    6                   X  +
                           |
                    XXXXX  |
                           |
                           |
    5                      +
                           |
                           |
                  XXXXXXX  |
                           |
    4                     S+
                        X  |
                     XXXX  |
                           |
                           |
    3             XXXXXXX  +
                           |
                           |T
                     XXXX  |
                           |
    2                     M+  SE6: Solve communication problems in English
                  XXXXXXX  |
                           |
                    XXXXX  |
                           |S SE8: Speak English only for whole lesson
    1                      +  SE1: Express opinion in English
                     XXXX  |
                           |
                     XXXX  |
                           |
    0                  XX S+M
                           |
                     XXXX  |
                           |  SE4: Speak English during group discussions
                       XX  |  SE7: Ask questions in English
   -1                      +
                           |S
                           |  SE2: Use phrases and skills that I learn
                           |  SE3: Speak English during pair work
                           |
   -2                   X T+
                           |
                       XX  |
                           |T
                           |
   -3                 |    +

Less self-efficacy (persons) Easier to endorse (items)
--------------------------------------------------------------------------------
```

Figure 7. Wright Map Showing Speaking Self-Efficacy Item-Person Responses

Note. M = mean, S = one standard deviation, T = two standard deviations, X = one student, *N* = 60. Measure values are in Rasch logits.

I had not anticipated that SE2 would also be easy to endorse, although on reflection perhaps the fact that the participants were full-time university students instilled in them the confidence to be able to use what they were taught.

Table 16. *Rasch Item Statistics for the Speaking Self-Efficacy Items*

Item	Measure	Model *SE*	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt-measure correlation
SE6	1.97	0.23	0.95	-0.20	0.90	-0.50	.75
SE8	1.22	0.24	1.21	1.10	1.29	1.50	.75
SE1	1.10	0.24	0.72	-1.60	0.75	-1.40	.79
SE4	-0.68	0.27	0.93	-0.30	0.79	-0.80	.77
SE7	-0.75	0.27	1.32	1.60	1.12	0.50	.57
SE2	-1.30	0.29	0.97	-0.10	0.95	0.00	.69
SE3	-1.55	0.29	0.86	-0.60	0.68	-0.90	.74

Note. All statistics are based on Rasch logits.

Dimensionality

Regarding dimensionality, the Rasch PCA of item residuals showed that 62.60% of the total variance was explained by the Rasch model, with an eigenvalue of 11.72. An additional 10.40% of the variance was explained by the first residual contrast, with an eigenvalue of 1.95, which is just within the level that can be attributed to random noise (Linacre, 2017). Table 17 shows the loading of item residuals from the speaking self-efficacy questionnaire. Items SE3 and SE4 had positive loadings above .40, while items SE6 and SE7 had negative loadings below -.40. Therefore, these items should be examined further to ensure that they are not tapping into unrelated dimensions. In the positive case, the two items specifically mention speaking in pairs and groups, so they could be addressing a subcomponent of free oral production. It is difficult to imagine a distinct relationship between the two negative items—speaking English for a whole lesson and asking questions—or any of the remaining items (SE1, SE2, and SE6), although overall the model appears to be functioning unidimensionally to a satisfactory degree.

Table 17. *Rasch PCA of Item Residuals for the Self-Efficacy Items*

Item	Loading	Measure	Infit MNSQ	Outfit MNSQ
SE3	.79	-1.55	0.86	0.68
SE4	.54	-0.68	0.93	0.79
SE2	.34	-1.30	0.97	0.95
SE1	.33	1.10	0.72	0.75
SE6	-.28	1.97	0.95	0.90
SE8	-.59	1.22	1.21	1.29
SE7	-.62	-0.75	1.32	1.12

Note. All statistics are based on Rasch logits.

Person Measure Descriptive Statistics

Descriptive statistics for the speaking self-efficacy person measures were also examined to determine if the results were normally distributed. The results for the combined data set, presented in Rasch logits, are displayed in the second column of Table 18, followed by the subsets of the data collected each time the questionnaire was administered. Levels of skewness and kurtosis were all within acceptable ranges. Furthermore, converted *z*-scores of the person measures all fell within the range of -3.29 to 3.29, confirming that the results for the speaking self-efficacy measure were normally distributed with no outliers.

Table 18. *Descriptive Statistics for Self-Efficacy Person Measures*

	Combined	Time 1	Time 2	Time 3
M measure	2.14	1.29	2.43	2.70
M SE	0.28	0.43	0.51	0.45
95% CI	[1.59, 2.69]	[0.39, 2.19]	[1.35, 3.40]	[1.77, 3.63]
SD	2.13	1.92	2.29	1.99
Skewness	-0.05	-0.33	-0.05	-0.03
SES	0.31	0.51	0.51	0.51
Kurtosis	-0.42	-0.67	-0.22	-0.89
SEK	0.61	0.99	0.99	0.99

Note. N = 60 for the combined data set; n = 20 for the Time 1, Time 2, and Time 3 data subsets. All statistics are based on Rasch logits.

Rasch Analysis of the Speaking Anxiety Questionnaire

The speaking anxiety questionnaire (Appendix M) consisted of six items, which was administered three times. These measures were again entered into a single analysis using stacked participant codes, resulting in three rows of data for each participant. The 20 participants thus created a data set of 360 individual item responses, as no data points were missing. This analysis was then examined for model fit and category functioning. Once again, a six-point Likert scale was used, whereby 1 = *Strongly disagree*, 2 = *Disagree*, 3 = *Moderately disagree*, 4 = *Moderately agree*, 5 = *Agree*, and 6 = *Strongly agree*. In this case, all categories were found to be functioning satisfactorily, with the minimum number of observed counts being 18 (strongly disagree) and the separation between each Andrich threshold surpassing the minimum acceptable value of 0.59 (Elwood, 2011). The results of the six-point scale are shown in Table 19 and the corresponding category probability curves are shown in Figure 8.

Table 19. *Category Structure Functioning for Speaking Anxiety*

Category	Observed count (%)	Average measure	Outfit MNSQ	Andrich threshold	Threshold change
1. Strongly disagree	18 (5)	-2.73	1.05	—	—
2. Disagree	54 (15)	-1.54	1.10	-3.18	—
3. Moderately disagree	87 (24)	-0.35	1.05	-1.42	1.76
4. Moderately agree	89 (25)	0.43	0.91	0.02	1.44
5. Agree	91 (25)	1.53	0.92	0.97	0.95
6. Strongly agree	21 (6)	3.00	1.01	3.61	2.64

Note. All statistics are based on Rasch logits.

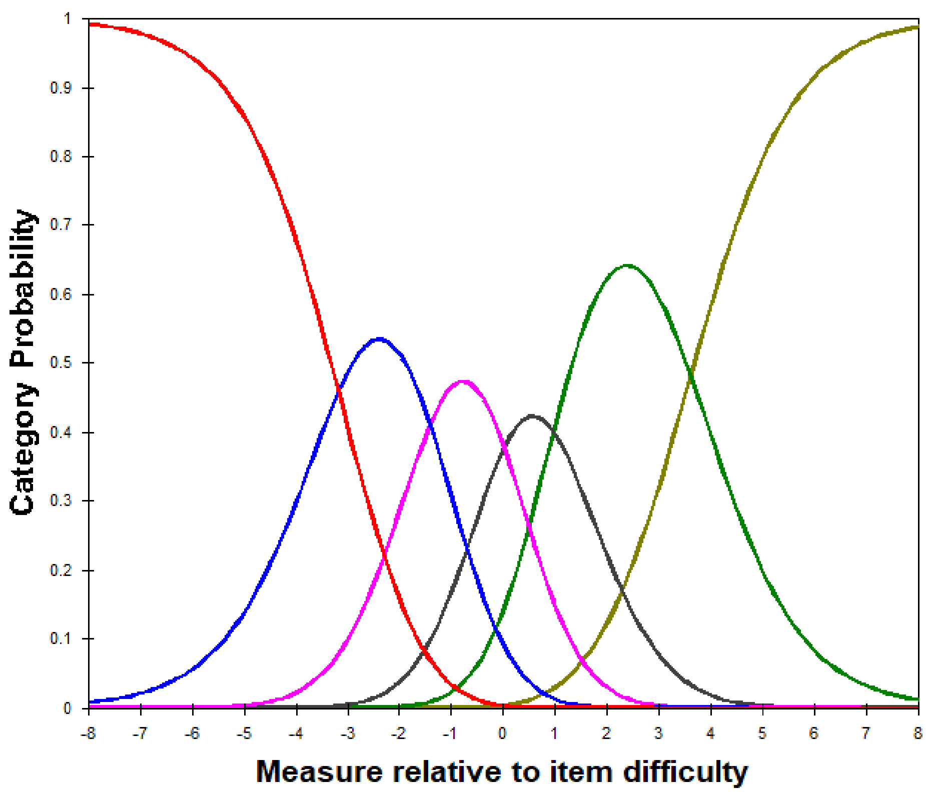

Figure 8. Rating Scale Performance of Speaking Anxiety Measured on a 6-Point Likert Scale

Item Fit and Endorsability

Table 20 shows the fit statistics for the eight questionnaire items, revealing that all items were within the acceptable infit and outfit MNSQ ranges of 0.50 to 1.50.

Furthermore, all ZSTD infit and outfit values are below 2.00, indicating compatibility with the Rasch model. Finally, all point-measure correlations were well above .40, which signified consistency of polarity and a strong relationship among these items.

Table 20. *Rasch Item Statistics for the Speaking Anxiety Items*

Item	Measure	Model *SE*	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt-measure correlation
ANX1	0.46	0.16	0.88	-0.64	0.92	-0.37	.81
ANX2	0.35	0.16	0.98	-0.07	1.02	0.17	.79
ANX4	0.32	0.16	1.03	0.23	1.00	0.03	.79
ANX3	-0.08	0.17	0.95	-0.25	0.95	-0.20	.81
ANX5	-0.25	0.17	1.21	1.16	1.18	1.01	.73
ANX6	-0.79	0.17	0.89	-0.54	0.93	-0.32	.76

Note. All statistics are based on Rasch logits.

The average person measure was 0.21, indicating that the items were well targeted at the participants, which is confirmed visually by the Wright map (Figure 9). The Rasch reliability estimate was .85 for both items and persons, while item separation was 2.36 and person separation was 2.39. According to Fisher's (2007) scale, these results indicate good reliability and a fair degree of separation.

Regarding the hierarchy of difficulty, *I feel nervous when speaking English to partner in pair work activities* (ANX1), proved to be the most difficult item to endorse (i.e., the least anxiety-inducing), followed by ANX2, *I feel nervous about participating in English group discussions,* and ANX4, *I am worried that other students think that my English speaking skills are bad*. I had correctly hypothesized that item 1 would not be anxiety-inducing, as it is the least face-threatening communication context, although I had expected items 2 and 4 to be more anxiety-inducing.

```
-------------------------------------------------------------------------------------
More anxiety (persons) More difficult to endorse (items)|

                X  |
                   |
   3               +
                X  |
                   |
                   |
             XXXX  |
                   |
   2           XX  +
                  S|
               XX  |
                   |
             XXXX  |
            XXXXX  |
   1               +
              XXX  |T
            XXXXX  |
                   |S ANX1: Nervous in pair work
             XXXX  |  ANX2: Nervous in group discussions ANX4: Others think my speaking is bad
               XX M|
   0               +M
                X  |  ANX3: Other students speak better
             XXXX  |  ANX5: Worried about making mistakes
             XXXX  |S
                   |
               XX  |T ANX6: Nervous when someone doesn't understand me
  -1          XXX  +
                   |
               XX  |
              XXX S|
                   |
               XX  |
  -2               +
               XX  |
                   |
                   |
                X  |
                   |
  -3               +
                  T|
                   |
                X  |
                   |
                   |
  -4               +
                X  |

Less anxiety (persons) Easier to endorse (items)
-------------------------------------------------------------------------------------
```

Figure 9. Wright Map Showing Anxiety Item-Person Responses
Note. M = mean, S = one standard deviation, T = two standard deviations, X = one student, *N* = 60 (one student response did not fit on map). Measure values are in Rasch logits.

This result suggests that the students felt relatively relaxed when interacting with each other, even in group settings, and that they were not overly concerned about how their speaking ability was perceived by their classmates. In other words, it implies that they felt comfortable in each other's company during lessons.

The easiest item to endorse was ANX6, *I feel nervous when someone doesn't understand what I say in English*, followed by ANX5, *I am worried about making mistakes when speaking English*. I had correctly hypothesized that these items would induce more anxiety, on the basis that communication problems are likely to cause stress—especially in a foreign language—and that accuracy tends to be heavily prioritized in pre-university formal English learning in Japan.

Dimensionality

Regarding dimensionality, the Rasch PCA of item residuals showed that 64.10% of the total variance was explained by the model, with an eigenvalue of 10.72. An additional 14.60% of the variance was explained by the first residual contrast, with an eigenvalue of 2.45. Although this value is above 2.00—the level that can be attributed to random noise—it is less than the practical criterion of 3.00 mentioned by Linacre (2017). Therefore, the construct can be considered sufficiently unidimensional for Rasch analysis purposes. Table 21 shows the loading of item residuals from the speaking anxiety questionnaire. Items ANX1 and ANX2 had high positive loadings, while items ANX4 and ANX5 had negative loadings below -.40. The two positive cases specifically mention speaking in pairs and groups so, as with self-efficacy, they could be addressing a subcomponent of free oral production. Again, it is less easy to imagine a separate relationship between the two negative items—anxiety about other students speaking

English better and about making mistakes—or the remaining items (ANX3, and ANX6), although in general the model seems to be functioning unidimensionally to a satisfactory degree.

Table 21. *Rasch PCA of Item Residuals for the Anxiety Items*

Item	Loading	Measure	Infit MNSQ	Outfit MNSQ
ANX1	.86	0.46	0.88	0.92
ANX2	.81	0.35	0.98	1.02
ANX3	-.05	-0.08	0.95	0.95
ANX6	-.12	-0.79	0.89	0.93
ANX5	-.68	-0.25	1.21	1.18
ANX4	-.76	0.32	1.03	1.00

Note. All statistics are based on Rasch logits.

Person Measure Descriptive Statistics

Finally, descriptive statistics for the speaking anxiety person measures were examined to determine if the results were normally distributed. Results for the combined data set are displayed in the second column of Table 22, followed by the subsets of the data collected each time the questionnaire was administered. Levels of skewness and kurtosis were all within acceptable ranges and converted *z*-scores of the person measures all fell within the range of -3.29 to 3.29.

Table 22. *Descriptive Statistics for Anxiety Person Measures*

	Combined	Time 1	Time 2	Time 3
M measure	0.21	0.56	0.16	-0.07
M SE	0.22	0.28	0.37	0.46
95% CI	[-0.22, 0.65]	[-.03, 1.14]	[-.62, .94]	[-1.04, 0.89]
SD	1.68	1.24	1.66	2.07
Skewness	0.06	-0.11	-0.55	0.73
SES	0.31	0.51	0.51	0.51
Kurtosis	0.83	-0.35	1.21	1.05
SEK	0.61	0.99	0.99	0.99

Note. N = 60 for the combined data set; *n* = 20 for the Time 1, Time 2, and Time 3 data subsets. All statistics are based on Rasch logits.

Rasch Analysis of the Rating Scales

A Rasch FACETS analysis was conducted to confirm the validity of the Fluency and Communicative Adequacy rating scales, and to produce logit measures of each participant's rated performance. The three raters had been shown selected video clips of each student participant in a randomized order, so that they were not aware which data collection point the clip was taken from. Each rater observed all 20 participants during their three discussions, amounting to 180 data points for each scale with no missing cases. As mentioned previously, the following three facets were entered into the model: person ability (i.e., based on the scores assigned by the raters); rater severity; and task difficulty (i.e., the three different discussion topics).

Perceived Fluency

A five-point numeric scale was used during the rating session (see Appendix R). However, the minimum of 10 observations was not met for the lowest category, as a score of 1—with the descriptors *Speaks very slowly* and *Frequent and lengthy hesitation at various points*—was only assigned four times. Therefore, the two lowest categories were collapsed into one, which was labeled *Disfluent speech* in Table 23. The results of this reduced four-point scale show that the minimum of 10 observations per category was now met. Furthermore, the average measure increased in accordance with the rating scale categories, the Andrich thresholds advanced, and separation between categories was below 5.00 in each case. The outfit mean square statistic was also below 2.00 for all categories. The corresponding category probability curves are shown in Figure 10, demonstrating how the rating scale conforms to the model expectations, as each scoring category has its own distinct peak. After examining these data, it was concluded that this adapted rating scale was performing in a satisfactory manner.

Table 23. *Category Structure Functioning for Perceived Fluency*

Category	Observed count	Average measure	Outfit MNSQ	Andrich threshold	Threshold change
2. Less fluent speech	56	-4.75	.71	—	—
3. —	71	-2.44	1.02	-4.48	—
4. —	41	1.65	.81	.10	4.58
5. More fluent speech	12	3.61	1.09	4.38	4.28

Note. All statistics are based on Rasch logits.

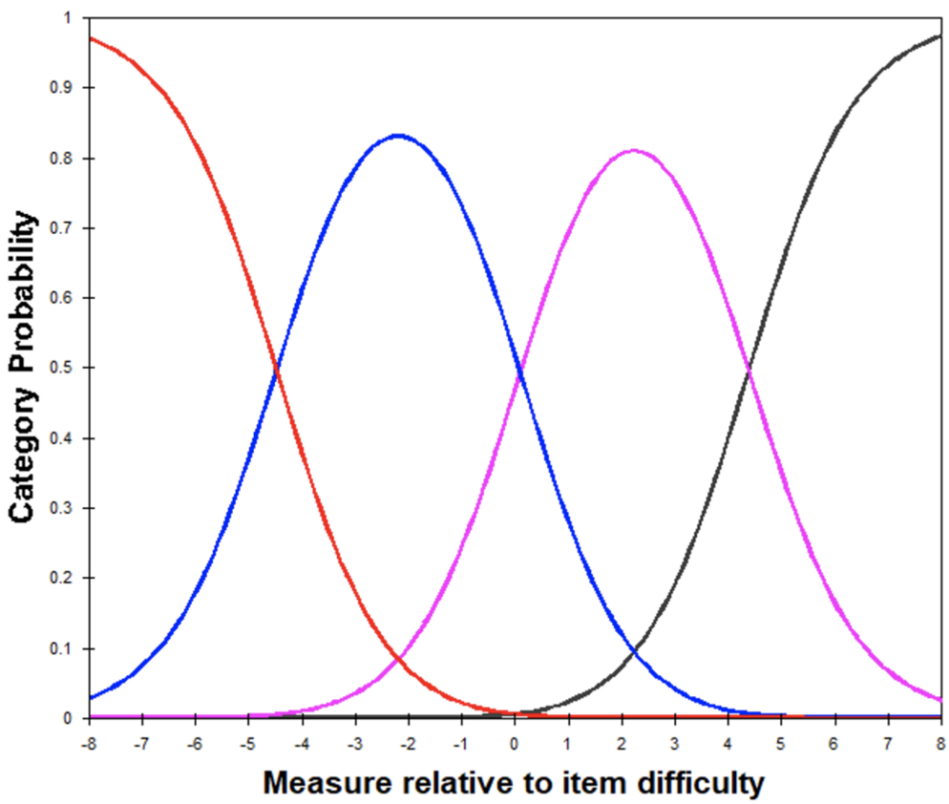

Figure 10. Rating Scale Performance of Perceived Fluency Measured on a 4-Point Scale

The Wright map (Figure 11) shows the Rasch estimates of the perceived levels of fluency, centered around a mean of approximately -2.00 logits. The first column shows a logit scale ranging from -6.00 to +5.00, indicating a wide spread of student abilities. The second column shows the actual distribution of abilities, with several students receiving the minimum possible score at one or more time points and one receiving the maximum. The participant separation statistic was 2.40 with a reliability estimate of .85, which is considered fair and good, respectively, according to Fisher's (2007) criteria. The third column shows that the raters varied slightly in their severity by 1.20 logits, while the fourth column shows a similar difference between the topics (0.82 logits) in terms of the ratings produced. However, both statistics are much smaller than the logit spread for student ability (14.24), indicating relative consistency in terms of rater and topic effect.

Table 24 shows the results from the FACETS analysis for the raters, ranked by severity—as shown by the measure in logits—with the most severe rater at the top. The fit statistics show that all the raters fit the model, thus demonstrating sufficient internal consistency, as all infit mean squares were well within the acceptable range of 0.50 to 1.50 that was set for the current study (Linacre, 2002b). Furthermore, the number of exact agreements among the raters was 56.10%, which is comfortably above Wolfe and Smith's (2007) recommended criterion of 40.00%. Nevertheless, the fixed chi-square value of 8.30 with 2 degrees of freedom was significant at $p = .02$, confirming the existence of statistical differences in rater severity.

```
+-----------------------------------------------------+
|Measure| +Student| -Rater |      -Topic      |Scale|
|-------+---------+--------+------------------+-----|
|   5   + *       +        +                  + (5) |
|       | *       |        |                  |     |
|       |         |        |                  | --- |
|       |         |        |                  |     |
|   4   +         +        +                  +     |
|       | ***     |        |                  |     |
|       |         |        |                  |     |
|       | **      |        |                  |     |
|   3   +         +        +                  +     |
|       | *       |        |                  |     |
|       |         |        |                  |     |
|       | *       |        |                  |  4  |
|   2   +         +        +                  +     |
|       | ***     |        |                  |     |
|       |         |        |                  |     |
|       | **      |        |                  |     |
|   1   +         +        +                  +     |
|       | **      |        |                  |     |
|       |         | Rater1 | Favorite season  |     |
|       | *       |        |                  |     |
*   0   *         * Rater2 * Favorite place   * --- *
|       | ***     |        |                  |     |
|       | *       | Rater3 | School life      |     |
|       |         |        |                  |     |
|  -1   + **      +        +                  +     |
|       |         |        |                  |     |
|       |         |        |                  |     |
|       | **      |        |                  |     |
|  -2   +         +        +                  +     |
|       | ****    |        |                  |  3  |
|       | ***     |        |                  |     |
|       |         |        |                  |     |
|  -3   +         +        +                  +     |
|       |         |        |                  |     |
|       | *****   |        |                  |     |
|       | **      |        |                  |     |
|  -4   +         +        +                  +     |
|       | ***     |        |                  |     |
|       |         |        |                  | --- |
|       | ***     |        |                  |     |
|  -5   +         +        +                  +     |
|       | *****   |        |                  |     |
|       | ***     |        |                  |     |
|       |         |        |                  |     |
|  -6   + ******* +        +                  + (2) |
|-------+---------+--------+------------------+-----|
```

Figure 11. Wright Map for the Perceived Fluency Rating Scale
Note. * = one student, *N* = 60. Measure values are in Rasch logits.

Raters

Table 24 shows the results from the FACETS analysis for the raters, ranked by severity—as shown by the measure in logits—with the most severe rater at the top. The fit statistics show that all of the raters fit the model, thus demonstrating sufficient internal consistency, as all infit mean squares were well within the acceptable range of 0.50 to 1.50 that was set for the current study (Linacre, 2002b). Furthermore, the number of exact agreements among the raters was 56.10%, which is comfortably above Wolfe and Smith's (2007) recommended criterion of 40%. Nevertheless, the fixed chi-square value of 8.30 with 2 degrees of freedom was significant at $p = .02$, confirming the existence of statistical differences in rater severity.

Table 24. *Rater Statistics for Perceived Fluency*

Rater	Total count	Fair average	Measure	Model *SE*	Infit MNSQ	Outfit MNSQ
1	60	3.01	0.60	0.30	0.87	0.82
2	60	3.11	-0.01	0.29	0.98	0.91
3	60	3.23	-0.60	0.29	1.04	1.02

Note. All statistics are based on Rasch logits. Separation = 1.80; reliability = .76.

Descriptive Statistics

Following Rasch analysis, descriptive statistics for the perceived fluency ratings were examined. The results for the combined data set, presented in Rasch logits, are displayed in the second column of Table 25, followed by the subsets of the ratings for each data collection point. Levels of skewness and kurtosis were all within acceptable ranges, although the kurtosis statistics for Times 2 and 3 were approaching the upper and lower bounds, respectively. Converted *z*-scores of the student participant measures all fell within the range of -3.29 to 3.29.

Table 25. *Descriptive Statistics for Perceived Fluency Ratings*

	Combined	Time 1	Time 2	Time 3
M measure	-1.99	-2.84	-2.34	-0.80
M SE	0.45	0.66	0.78	0.86
95% CI	[-2.90, -1.09]	[-4.22, -1.46]	[-3.96, -0.72]	[-2.60, 1.01]
SD	3.50	2.95	3.46	3.85
Skewness	0.50	-0.60	0.97	-0.11
SES	0.31	0.51	0.51	0.51
Kurtosis	-0.60	-0.52	1.61	-1.59
SEK	0.61	0.99	0.99	0.99

Note. N = 180 for the combined data set; n = 60 for the Time 1, Time 2, and Time 3 data subsets. All statistics are based on Rasch logits.

Communicative Adequacy

A five-point numeric scale was also used for Communicative Adequacy, which was rated concurrently with Fluency (see Appendix R). Once again, the minimum of 10 observations was not met for the lowest category, as a score of 1—with the descriptors *Contributions are often irrelevant* and *Contributions are difficult to understand*—was only assigned three times. Therefore, the two lowest categories were collapsed into one, which has been labeled *Inadequate speech* in Table 26. The results of the reduced four-point scale show that the minimum of 10 observations per category was now met. Furthermore, the average measure increased in accordance with the rating scale categories, the Andrich thresholds advanced, and separation between categories was below 5.00 in each case. The outfit mean square statistic was also below 2.00 for all categories. The corresponding category probability curves are shown in Figure 12, demonstrating how the rating scale conforms to the model expectations. It was thus concluded that this adapted rating scale was performing in a satisfactory manner.

Table 26. *Category Structure Functioning for Communicative Adequacy*

Category	Observed count	Average measure	Outfit MNSQ	Andrich threshold	Threshold change
2. Inadequate speech	54	-5.16	0.70	—	—
3. —	62	-1.87	1.00	-4.45	—
4. —	47	1.90	1.00	0.44	4.89
5. Adequate speech	17	3.35	1.20	4.00	3.56

Note. All statistics are based on Rasch logits.

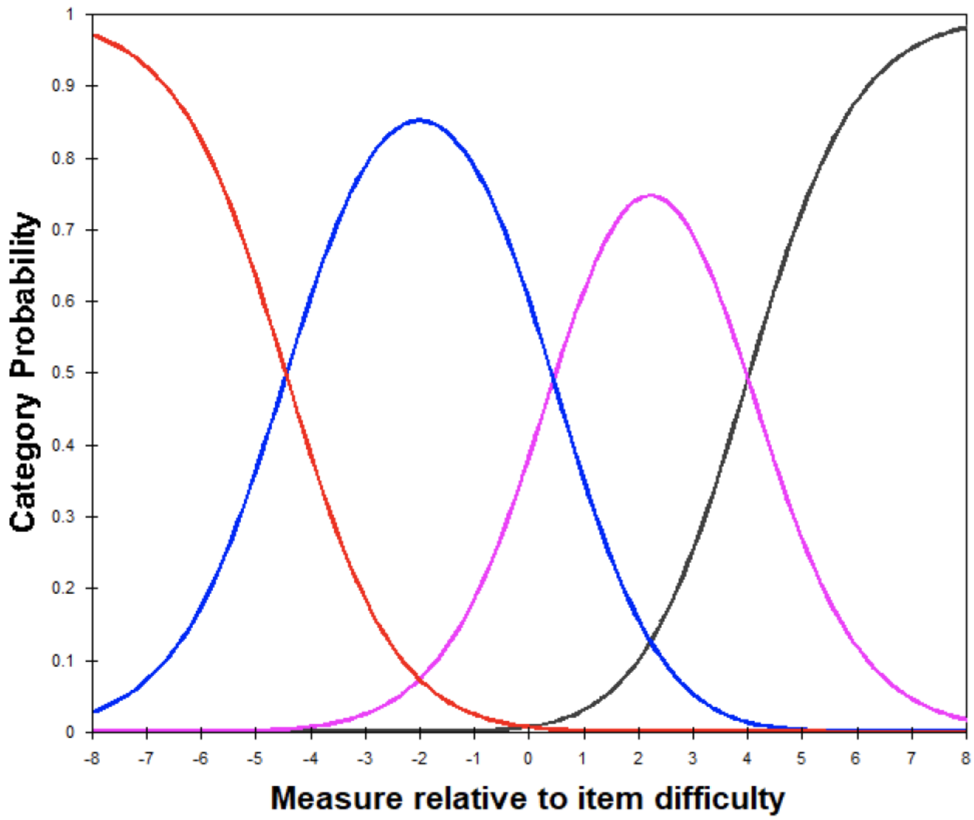

Figure 12. Rating Scale Performance of Communicative Adequacy Measured on a 4-Point Scale

```
+------------------------------------------------------------------------+
|Measure|  +Student  | -Rater |              -Topic              |Scale|
|----  -+------------+--------+----------------------------------+-----|
|   5   + *          +        +                                  + (5) |
|       | **         |        |                                  |     |
|       |            |        |                                  |     |
|       |            |        |                                  |     |
|   4   +            +        +                                  + --- |
|       |            |        |                                  |     |
|       |            |        |                                  |     |
|       | *****      |        |                                  |     |
|   3   +            +        +                                  +     |
|       |            |        |                                  |     |
|       | ****       |        |                                  |     |
|       |            |        |                                  |  4  |
|   2   + ****       +        +                                  +     |
|       |            |        |                                  |     |
|       |            |        |                                  |     |
|       | *          |        |                                  |     |
|   1   + **         + Rater1 +                                  +     |
|       | ***        |        |                                  |     |
|       |            |        |                                  | --- |
|       |            |        | Favorite season                  |     |
*   0   * **         * Rater2 * Favorite place   School life     *     *
|       |            |        |                                  |     |
|       | *****      |        |                                  |     |
|       |            | Rater3 |                                  |     |
|  -1   +            +        +                                  +     |
|       |            |        |                                  |     |
|       |            |        |                                  |     |
|       | *          |        |                                  |     |
|  -2   + *****      +        +                                  +  3  |
|       |            |        |                                  |     |
|       |            |        |                                  |     |
|       |            |        |                                  |     |
|  -3   +            +        +                                  +     |
|       |            |        |                                  |     |
|       | ***        |        |                                  |     |
|       | **         |        |                                  |     |
|  -4   +            +        +                                  +     |
|       |            |        |                                  |     |
|       |            |        |                                  | --- |
|       |            |        |                                  |     |
|  -5   + ****       +        +                                  +     |
|       | *********  |        |                                  |     |
|       |            |        |                                  |     |
|       |            |        |                                  |     |
|  -6   + *******    +        +                                  + (2) |
|-------+------------+--------+----------------------------------+-----|
```

Figure 13. Wright Map for the Communicative Adequacy Rating Scale
Note. * = one student, *N* = 60. Measure values are in Rasch logits.

The Wright map (Figure 13) shows the Rasch estimates of the perceived levels of communicative adequacy, centered around a mean of approximately -1.50 logits. As with the perceived fluency scale, the first column shows a logit scale ranging from -6.00 to +5.00, indicating a wide spread of student abilities. Again, several students received the minimum possible score at one or more time points with only one receiving the maximum. The participant separation ratio was 2.58 with a reliability estimate of .87, which is considered fair and good, respectively, according to Fisher's (2007) criteria. The raters varied slightly more in their severity (1.74 logits) than when rating fluency, although the same order of severity was maintained. However, there was little difference between the topics (0.31 logits) in terms of the difficulty ratings produced. Again, both statistics are much smaller than the logit spread for student ability (13.24), indicating relative consistency in terms of rater and topic effect.

Raters

Table 27 shows the results from the FACETS analysis for the raters, ranked by severity with the most severe rater at the top. The fixed chi-square value of 17.80 with 2 degrees of freedom was significant at $p < .001$, confirming the statistical differences in rater severity. Nevertheless, all of the raters fit the model, thus demonstrating sufficient internal consistency, as all infit mean squares were well within the acceptable range of 0.50 to 1.50. Furthermore, the number of exact agreements among the raters was 49.40%, which is above Wolfe and Smith's (2007) recommended criterion of 40.00%.

Table 27. *Rater Statistics for Communicative Adequacy*

Rater	Total count	Fair average	Measure	Model *SE*	Infit MNSQ	Outfit MNSQ
1	60	3.02	0.92	0.30	0.92	0.86
2	60	3.19	-0.10	0.29	0.96	0.96
3	60	3.35	-0.82	0.28	1.09	1.08

Note. All statistics are based on Rasch logits. Separation = 2.85; reliability = .89.

Descriptive Statistics

Following Rasch analysis, descriptive statistics for the communicative adequacy ratings were examined. The results for the combined data set, presented in Rasch logits, are displayed in the second column of Table 28, followed by the subsets of the ratings for each data collection point. Levels of skewness and kurtosis were all within acceptable ranges, although the kurtosis statistics for the combined and Time 1 data sets (-1.26 and -1.14) indicate that they were trending towards a platykurtic distribution. Converted *z*-scores of the student participant measures all fell within the range of -3.29 to 3.29.

Table 28. *Descriptive Statistics for Communicative Adequacy Ratings*

	Combined	Time 1	Time 2	Time 3
M measure	-1.53	-3.14	-2.19	0.75
M SE	0.50	0.79	0.74	0.73
95% CI	[-2.49, -0.57]	[-4.80, -1.48]	[-3.74, -.65]	[-0.78, 2.28]
SD	3.71	3.54	3.30	3.27
Skewness	0.14	-0.59	0.61	-0.65
SES	0.31	0.51	0.51	0.51
Kurtosis	-1.26	-1.14	-0.72	0.00
SEK	0.61	0.99	0.99	0.99

Note. N = 180 for the combined data set; n = 60 for the Time 1, Time 2, and Time 3 data subsets. All statistics are based on Rasch logits.

Summary of Preliminary Analyses

In this chapter, I outlined how the quantitative questionnaire and rating scale data used in this study were examined for reliability and validity. Specifically, the dimensionality as well as model and item fit of the extraversion, speaking self-efficacy, and speaking anxiety measures obtained from the questionnaire responses were described. Next, the descriptive statistics for the participant measures gathered from these analyses were summarized. A similar process was followed regarding the rating scale data, as the model fit and reliability of the perceived fluency and communicative adequacy ratings were described. Fit and functioning of the three facets entered into the model—students, raters, and topics—were also investigated. Finally, descriptive statistics of the rating measures were presented. Once category functioning and item or rater fit had been established, all the questionnaire and rating scale data met the necessary criteria for use in this study.

CHAPTER 5

RESULTS

In this chapter, I present the results of the analyses that were conducted to answer the research questions (stated in Chapter 2). Each research question is presented in turn, followed by an explanation of the relevant statistical analyses and their results. All statistical analyses were conducted using JASP version 0.19.1. Regarding effect sizes, Plonsky and Oswald's (2014) benchmark guidelines for L2 research were adopted in this study. The relevant Cohen's *d* values for within-groups mean difference contrasts are 0.60, 1.00, and 1.40 for small, medium, and large effect sizes, respectively. For correlation coefficients, Pearson's *r* values of approximately .25 are considered small, .40 medium, and .60 large. These guidelines are considerably more stringent than Cohen's (1988) original benchmarks, which were intended for psychological research and have been found to underestimate the magnitude of effects typically obtained in L2 research (Plonsky & Oswald, 2014). A fuller discussion and interpretation of these results, with reference to existing research and foreign language teaching pedagogy, is provided in Chapter 6.

Research Question 1: The Extent of Speaking Fluency Development

Research question 1 investigated the extent to which the student participants' speaking fluency developed during a one-semester course of spoken interaction. The question was subdivided into three components reflecting the various procedures used to measure different aspects of fluency in monologue and group discussion activities, as well as ratings of perceived fluency. Each of these subsections is presented below.

Monologue Activities

The participants' utterance fluency was measured on an individual basis using monologue activities conducted at the three measurement points during the semester. An overall measure of pruned speech rate was calculated by removing all disfluent syllables from the speech sample and dividing the remaining number of syllables by the total time. Specific measures of speed (i.e., rate of articulation), breakdown (i.e., pausing phenomena), and repair (i.e., number of repetitions and self-corrections) were also calculated. Refer to Table 4 in Chapter 3 for a full list of the utterance measures included. One-way repeated measures ANOVAs were conducted to examine the extent of any changes regarding these measures.

Pruned Speech Rate

Taking the overall measure of pruned speech rate first, the descriptive statistics (see Table 29) show an increase in mean speech rate of 5.47 syllables between Times 1 and 2, and 9.73 syllables between Times 2 and 3. Values for skewness and kurtosis were within the acceptable limits of -2.00 to +2.00 (Hair et al., 2022). Also, the data were normally distributed for each time point, as assessed by the Shapiro-Wilk test (p = .40, .12, and .61, respectively). There was one outlier at Time 3, with this participant outperforming the next highest score by 16 syllables per minute. However, this participant also had the highest score at Time 1, and the second highest at Time 2, so it was not considered an especially unlikely result. This case exemplifies the fact that data based on real speech can produce results that are unexpected but not unlikely in individual cases. All participants were observed to be on task during all recorded activities; therefore, this and all subsequently mentioned outliers have been retained in the analysis. Mauchly's test indicated that the assumption of sphericity was met, with non-significant results for the Chi-square test $\chi^2(2) = 3.39$, $p = .18$.

Table 29. *Descriptive Statistics for Pruned Speech Rate*

	Time 1	Time 2	Time 3
M	79.90	85.37	95.10
SE	4.20	4.74	5.31
95% CI	[71.11, 88.69]	[75.44, 95.30]	[83.99, 106.21]
SD	18.78	21.21	23.74
Skewness	0.69	0.46	0.46
SES	0.51	0.51	0.51
Kurtosis	-0.02	-0.76	0.65
SEK	0.99	0.99	0.99

Note. $N = 20$ at each time point.

The repeated measures ANOVA revealed statistically significant changes in pruned speech rate over time, $F(2, 38) = 7.69$, $p = .002$, $\eta^2 = .29$, with 29.00% of the variance accounted for by the model. Post hoc *t*-tests with a Bonferroni adjustment, to keep the Type I error at 5.00% overall, revealed a significant pairwise difference between Time 1 and Time 3 only ($p = .001$, $d = 0.71$) but not between Time 1 and Time 2 ($p = .52$, $d = 0.26$), or between Time 2 and Time 3 ($p = .053$, $d = 0.46$). However, as the latter is a marginal result—and has a larger effect size—it can be concluded that more gains in pruned speech rate occurred between Times 2 and 3, that is, in the second half of the semester.

Specific Utterance Fluency Measures

Before the specific utterance fluency measures were analyzed, bivariate correlation analysis was conducted to investigate potential multicollinearity. Although many significant correlations were identified, Tables 30 to 32 confirm that no variables correlated above the threshold of .80 at any of the three time points (Field, 2024); consequently, no variables were removed from the analysis. No clear pattern emerged regarding correlations between these specific measures and the composite measure of pruned speech rate. For example, it correlated most strongly with rate of mid-clause silent pauses (positive) at Time 1 (r = .54), mean duration of mid-clause silent pauses (negative) at Time 2 (r = -.62), and mean syllable duration (negative) at Time 3 (r = -.45).

Table 30. *Bivariate Correlations Between Utterance Fluency Variables (Time 1)*

	1	2	3	4	5	6	7	8
1. Syllable duration	—							
2. SP rate (end)	.40	—						
3. SP rate (mid)	-.12	.26	—					
4. FP rate (end)	-.35	-.75**	-.34	—				
5. FP rate (mid)	-.38	-.64**	-.48*	.60**	—			
6. SP mean (end)	.40	.60*	-.10	-.58**	-.39	—		
7. SP mean (mid)	.26	.34	.08	-.50*	-.69**	.54*	—	
8. Repetitions	-.23	-.59**	-.09	.67**	.50*	-.49*	-.39	—
9. Repairs	.03	-.24	.28	.09	.27	-.10	-.27	.29

Note. * p < .05, ** p < .01. SP = silent pause, FP = filled pause, mid = mid-clause, end = end-clause.

Table 31. *Bivariate Correlations Between Utterance Fluency Variables (Time 2)*

	1	2	3	4	5	6	7	8
1. Syllable duration	—							
2. SP rate (end)	-.27	—						
3. SP rate (mid)	.19	.19	—					
4. FP rate (end)	-.34	-.22	-.41	—				
5. FP rate (mid)	-.13	-.28	-.51*	.50*	—			
6. SP mean (end)	.11	.14	.32	-.59**	-.69**	—		
7. SP mean (mid)	.37	-.25	.46*	-.66**	-.59**	.68**	—	
8. Repetitions	.16	-.17	-.01	.03	.55*	-.28	-.03	—
9. Repairs	-.03	-.03	.63**	-.11	-.12	-.06	.14	.17

Note. * p < .05, ** p < .01. SP = silent pause, FP = filled pause, mid = mid-clause, end = end-clause.

Table 32. *Bivariate Correlations Between Utterance Fluency Variables (Time 3)*

	1	2	3	4	5	6	7	8
1. Syllable duration	—							
2. SP rate (end)	.22	—						
3. SP rate (mid)	.07	.42	—					
4. FP rate (end)	-.40	-.67**	-.52*	—				
5. FP rate (mid)	-.45*	-.51**	-.68**	.68**	—			
6. SP mean (end)	.25	.43	.36	-.49*	-.67*	—		
7. SP mean (mid)	.22	.12	.16	-.47*	-.47*	.65**	—	
8. Repetitions	.13	-.29	-.06	.18	.24	-.16	-.06	—
9. Repairs	-.33	-.26	-.04	.32	.46*	-.66**	-.61**	.01

Note. * $p < .05$, ** $p < .01$. SP = silent pause, FP = filled pause, mid = mid-clause, end = end-clause.

Having established that there were no redundant variables, repeated measures ANOVAs were conducted for each of the nine specific utterance fluency variables. The results are summarized in Table 33, with mean values at each time point presented alongside *p*-values and effect sizes of the associated ANOVA. Statistically significant results were produced for the following three variables: mean duration of end-clause silent pauses, mean duration of mid-clause silent pauses, and number of repairs per minute. Results are presented in detail for these measures only.

Table 33. *Summary of Results for Specific Utterance Fluency Variables*

	Time 1	Time 2	Time 3	*p*	η^2
Syllable duration	0.30	0.31	0.30	.26	.07
SP rate (end)	4.44	4.83	4.20	.41	.05
SP rate (mid)	12.01	13.12	12.70	.59	.03
FP rate (end)	2.07	2.10	2.43	.51	.04
FP rate (mid)	4.87	5.90	5.97	.21	.08
SP mean (end)	1.39	0.90	0.95	< .01*	.43
SP mean (mid)	0.92	0.80	0.77	< .01*	.22
Repetitions	2.83	3.47	3.17	.44	.04
Repairs	2.20	3.13	3.63	.02*	.19

Note. * Denotes statistically significant result. SP = silent pause, FP = filled pause, mid = mid-clause, end = end-clause.

Mean Duration of End-Clause Silent Pauses. Looking at the mean duration of end-clause silent pauses first, the descriptive statistics (see Table 34) show a relatively large decrease (0.49 seconds) in the mean duration of end-clause silent pauses between Times 1 and 2, and a slight increase—albeit with largely overlapping confidence intervals—between Times 2 and 3. The data were normally distributed at Times 1 and 3 as assessed by the Shapiro-Wilk test (p = .051 and .67, respectively). However, they were not normally distributed at Time 2 (p = .01), perhaps due to the presence of an outlier. Mauchly's test indicated that the assumption of sphericity was met, with non-significant results for the Chi-square test $\chi^2(2) = 5.73$, $p = .06$.

A repeated measures ANOVA showed significant changes over time, $F(2, 38) = 14.03$, $p < .001$, $\eta^2 = .43$, with 43.00% of the variance accounted for. Post hoc t-tests (Bonferroni adjusted) revealed a significant pairwise difference between Time 1 and Time 3 ($p < .001$, $d = 0.83$), indicating a small to medium effect size. However, the largest difference was between Time 1 and Time 2 ($p < .001$, $d = 0.93$), showing a medium effect size. There was no significant difference between Times 2 and 3 ($p = 1$, $d = 0.10$).

Table 34. *Descriptive Statistics for Mean Duration of Silent Pauses (End)*

	Time 1	Time 2	Time 3
M	1.39	0.90	0.95
SE	0.16	0.10	0.10
95% CI	[1.07, 1.72]	[.69, 1.10]	[.75, 1.15]
SD	0.70	0.43	0.43
Skewness	1.01	1.32	0.48
SES	0.51	0.51	0.51
Kurtosis	0.48	1.46	-0.41
SEK	0.99	0.99	0.99

Note. N = 20 at each time point.

Mean Duration of Mid-Clause Silent Pauses. Looking at mean duration of mid-clause silent pauses next, the descriptive statistics (see Table 35) show a decrease of .12 seconds between Times 1 and 2, which remained relatively stable between Times 2 and 3, with a decrease of only 0.03 seconds. The data were not normally distributed at Time 1 (p = .03), perhaps reflecting the presence of an outlier. However, the data were normally distributed at Times 2 and 3 (p = .06 and .27, respectively) with no outliers at these time points. Mauchly's test again revealed that the assumption of sphericity was met, with non-significant results for the Chi-square test $\chi^2(2) = 1.89$, $p = .39$.

Table 35. *Descriptive Statistics for Mean Duration of Silent Pauses (Mid)*

	Time 1	Time 2	Time 3
M	0.92	0.80	0.77
SE	0.07	0.05	0.04
95% CI	[.77, 1.07]	[0.69, 0.91]	[0.68, 0.86]
SD	0.33	0.24	0.19
Skewness	1.12	0.69	0.24
SES	0.51	0.51	0.51
Kurtosis	1.02	-0.43	-1.21
SEK	0.99	0.99	0.99

Note. N = 20 at each time point.

A repeated measures ANOVA showed significant changes over time, $F(2, 38) = 5.35$, $p = .009$, $\eta^2 = .22$, with the model accounting for 22.00% of the variance. Post hoc t-tests (Bonferroni adjusted) revealed a significant pairwise difference between Time 1 and Time 3 ($p = .01$, $d = 0.59$), indicating a small effect size, but not between Time 1 and Time 2 ($p = .06$, $d = 0.46$), or between Times 2 and 3 ($p = 1$, $d = 0.13$).

Number of Repairs per Minute. Regarding the number of repairs per minute, the descriptive statistics (see Table 36) show an increase in the number of repairs per minute of 0.93 between Times 1 and 2, and a further 0.5 between Times 2 and 3. The data were not normally distributed at Time 1 (p = .01), again perhaps due to the presence of an outlier. The data were normally distributed at Times 2 and 3 (p = .25 and .15, respectively) with no outliers at these time points. Mauchly's test showed that the assumption of sphericity was met, with non-significant results for the Chi-square test $\chi^2(2) = 5.62$, $p = .06$.

Table 36. *Descriptive Statistics for Number of Repairs per Minute*

	Time 1	Time 2	Time 3
M	2.20	3.13	3.63
SE	0.29	0.41	0.39
95% CI	[1.59, 2.81]	[2.27, 3.99]	[2.83, 4.44]
SD	1.30	1.84	1.73
Skewness	1.38	0.30	0.14
SES	0.51	0.51	0.51
Kurtosis	2.62	-0.96	-0.94
SEK	0.99	0.99	0.99

Note. N = 20 at each time point.

A repeated measures ANOVA showed significant changes over time, $F(2, 38)$ = 4.43, p = .02, η^2 = .19, with 19.00% of the variance accounted for. Post hoc t-tests (Bonferroni adjusted) revealed a significant pairwise difference between Time 1 and Time 3 only (p = .02, d = 0.88), indicating a small to medium effect size, but not between Time 1 and Time 2 (p = .19, d = 0.57), or between Time 2 and Time 3 (p = .94, d = 0.31), although the former showed a small effect size.

Group Discussions

Fluency in the recorded group discussions was measured both individually, in terms of the number of fluent syllables produced per minute (i.e., with filled pauses, repeated syllables, and false starts removed), and on a group basis, in terms of the mean length and number of gaps between turns. For the individual measure, a repeated measures ANOVA was conducted to examine the extent of any changes over the three measurement points. However, inferential statistics could not be used for the group-based measures as each group consisted of different members at each time point. Therefore, descriptive statistics only are presented for these measures.

Fluent Syllables Produced per Minute

The descriptive statistics (see Table 37) show an overall increase in the mean syllables produced per minute of 8.98, of which 7.73 occurred between Time 1 and Time 2. Values for skewness and kurtosis were within acceptable limits. The data were normally distributed for each time point, as assessed by the Shapiro-Wilk test (p = .98, .35, and .97, respectively), with one outlier at Time 1. Mauchly's test indicated that the assumption of sphericity was met, with non-significant results for the Chi-square test $\chi^2(2) = 0.76$, $p = .68$.

Table 37. *Descriptive Statistics for Fluent Syllables Produced per Minute*

	Time 1	Time 2	Time 3
M	30.75	38.48	39.73
SE	2.60	3.21	3.04
95% CI	[25.32, 36.19]	[31.76, 45.20]	[33.36, 46.09]
SD	11.61	14.36	13.60
Skewness	0.25	0.49	0.13
SES	0.51	0.51	0.51
Kurtosis	0.29	-0.69	-0.90
SEK	0.99	0.99	0.99

Note. N = 20 at each time point.

A repeated measures ANOVA revealed significant changes in the number of fluent syllables produced over time, $F(2, 38) = 14.84$, $p < .001$, $\eta^2 = .44$, with 44.00% of the variance accounted for. Post hoc t-tests (Bonferroni adjusted) revealed a significant pairwise difference between Time 1 and Time 3 (p = < .001, d = 0.68), and Time 1 and

Time 2 (p = < .001, d = 0.58), indicating a small effect size in both cases, but not between Times 2 and 3 (p = 1, d = 0.09).

Gaps Between Turns

Because a repeated measures ANOVA could not be used to analyze gaps between turns, these statistics are based on the average of all groups at each time point, as shown in Table 38. However, only groups that consisted of at least one of the 20 core participants—who were selected from the 40 total student participants—were included in the analysis. These groups, which consisted of different members at each time point, therefore included one, two, or three core participants. Any remaining places consisted of other class members, that is, those who were not selected as core participants. Due to these differing permutations, the number of groups used to calculate gaps between turns was not always identical at each time point.

Table 38. *Group Averages for Gaps Between Turns*

	Time 1	Time 2	Time 3
No. of gaps	33.67	21.75	23.55
Mean length	3.39	2.77	2.46
Total time	116.47	61.22	58.12

Note. N = 12 at Time 1 and Time 2, N = 11 at Time 3.

These results show that for each measure there was a large reduction between Time 1 and Time 2. Regarding the number of gaps and total time of gaps this reduction was close to 50.00%, with very little change observed between Time 2 and Time 3 in both cases. Regarding mean length of gaps, the reduction was 0.62 seconds between Times 1 and 2, and 0.31 seconds between Times 2 and 3. To summarize, the number and length of gaps between turns decreased by a large amount during the first half of the semester, resulting in a similarly large decrease in the total time of such gaps, which then stabilized in the second half of the semester.

Tables 39 to 41 present each measure individually. The number and total time of gaps between turns (Tables 39 and 41) exhibited wide confidence intervals and large standard deviations, reflecting a considerable amount of variance between groups. However, there was much less variance regarding the mean length of gaps from Time 2 onwards (Table 40). This finding was reflected not only in the narrower confidence intervals and standard deviations but also in the high kurtosis statistics, as many similar values produced a peaked or leptokurtic distribution.

Table 39. *Descriptive Statistics for Number of Gaps Between Turns*

	Time 1	Time 2	Time 3
M	33.67	21.75	23.55
SE	3.61	2.12	3.74
95% CI	[25.72, 41.62]	[17.09, 26.41]	[15.21, 31.88]
SD	12.51	7.39	12.41
Skewness	-0.05	-0.38	1.22
SES	0.64	0.64	0.66
Kurtosis	-1.18	-0.17	0.62
SEK	1.23	1.23	1.28

Note. *N* = 12 at Time 1 and Time 2; *N* = 11 at Time 3.

Table 40. *Descriptive Statistics for Mean Length of Gaps Between Turns*

	Time 1	Time 2	Time 3
M	3.39	2.77	2.46
SE	0.36	0.21	0.14
95% CI	[2.61, 4.17]	[2.31, 3.23]	[2.14, 2.77]
SD	1.23	0.72	0.47
Skewness	0.70	1.06	1.65
SES	0.64	0.64	0.66
Kurtosis	-0.74	3.34	4.06
SEK	1.23	1.23	1.28

Note. *N* = 12 at Time 1 and Time 2; *N* = 11 at Time 3.

Table 41. *Descriptive Statistics for Total Time of Gaps Between Turns*

	Time 1	Time 2	Time 3
M	116.47	61.22	58.12
SE	19.69	7.94	9.52
95% CI	[73.13, 159.82]	[43.75, 78.68]	[36.89, 79.34]
SD	68.22	27.50	31.59
Skewness	0.93	0.01	0.94
SES	0.64	0.64	0.66
Kurtosis	-0.43	-1.46	0.36
SEK	1.23	1.23	1.28

Note. N = 12 at Time 1 and Time 2; N = 11 at Time 3.

Native-Speaker Ratings

Each participant's performance during the group discussions was rated in terms of fluency and communicative adequacy (see Appendix R for the analytic rating descriptors). Although fluency is directly related to the research question, communicative adequacy was mainly included to check whether the demands of the task were being fulfilled. Logit measures, based on Rasch analysis, were used for the analysis—validation evidence for these measures was presented in Chapter 4. One-way repeated measures ANOVAs were conducted to investigate the extent of any changes over time. Table 42 summarizes the results, with mean values at each time point presented alongside *p*-values and effect sizes of the related ANOVA. Results are then presented in detail for each variable.

Table 42. *Summary of Results for Group Discussion Ratings*

	Time 1	Time 2	Time 3	p	η^2
Fluency ratings	-2.84	-2.34	-.80	.04	.16
CA ratings	-3.14	-2.19	.75	< .01	.50

Note. N = 60 at each time point. CA = communicative adequacy.
All statistics are based on Rasch logits.

Perceived Fluency

Taking perceived fluency first, the descriptive statistics (see Table 25 in Chapter 4) show an increase in the logit measures of 0.50 between Times 1 and 2, and 1.54 between Times 2 and 3. The Shapiro-Wilk test confirmed that the data were normally distributed for Times 1 and 2 (p = .18 and .24, respectively), although not at Time 3 (p = .049) due to a double peaked distribution towards the upper and lower ends. There was also one outlier at Time 2, who received the maximum fluency rating from all three raters. However, this participant also received the highest average rating at Time 1 and the joint-second highest at Time 3; therefore, this result was not deemed to be anomalous and was retained in the analysis. Mauchly's test indicated that the assumption of sphericity was met, with non-significant results for the Chi-square test $\chi^2(2) = 1.14$, p = .57.

The repeated measures ANOVA revealed significant changes in perceived fluency over time, $F(2, 38) = 3.64$, $p = .04$, $\eta^2 = .16$, albeit with only 16.00% of the variance accounted for. Post hoc t-tests (Bonferroni adjusted) showed a significant pairwise difference between Time 1 and Time 3 only ($p = .04$, $d = 0.59$), indicating a small effect size, but not between Time 1 and Time 2 ($p = 1$, $d = 0.14$), or between Time 2 and Time 3 ($p = .17$, $d = 0.49$).

Communicative Adequacy

Looking at communicative adequacy next, the descriptive statistics (see Table 28 in Chapter 4) showed an increase in the logit measures of .95 between Times 1 and 2, and 2.94 between Times 2 and 3. The Shapiro-Wilk test revealed that the data were normally distributed for Times 2 and 3 (p = .07, and .09, respectively), although the Time 1 data were not normally distributed (p = .02) due to a relatively large number of scores at the lower end. There were no outliers at any time point. Mauchly's test showed that the assumption of sphericity was met, with non-significant results for the Chi-square test $\chi^2(2) = 0.88$, $p = .32$.

The repeated measures ANOVA revealed significant changes in the raters' perceptions of communicative adequacy over time, $F(2, 38) = 19.19$, $p < .001$, $\eta^2 = .5$. In this case 50.00% of the variance was thus accounted for by the model—much more than was the case for perceived fluency. Post hoc *t*-tests (Bonferroni adjusted) showed a significant pairwise difference between Time 1 and Time 3 ($p < .001$, $d = 1.16$), indicating a medium to large effect size, and Time 2 and Time 3 ($p < .001$, $d = 0.87$), with a small to medium effect size, although not between Times 1 and 2 ($p = .46$, $d = 0.28$).

Correlations Between Fluency and Communicative Adequacy Measures

The three individual fluency measures—utterance fluency (i.e., represented by the composite measure of pruned speech rate), fluent syllables per minute, and perceived fluency—were entered into a bivariate correlation analysis, together with the communicative adequacy ratings. Utterance fluency was based on performance in the monologue activities, whereas syllables per minute and perceived fluency were based on performance in the group discussions. The analysis was conducted for each data collection time point. Tables 43 to 45 show that all measures were positively and significantly correlated at each point. Using Plonsky and Oswald's (2014) guidelines, large effect sizes were produced in almost every case, with Pearson's *r* ranging from .48 to .83.

Among the three fluency measures, the strongest correlation was consistently between utterance fluency and syllables per minute, followed by utterance and perceived fluency, and finally perceived fluency and syllables per minute. Communicative adequacy was most strongly associated with perceived fluency at each time point. Positive relationships were also observed between communicative adequacy and both utterance fluency and syllables per minute at each stage.

Table 43. *Bivariate Correlations Between Fluency and Communicative Adequacy Measures (Time 1)*

	1	2	3
1. Utterance fluency	—		
2. Fluent syllables per min.	.76***	—	
3. Perceived fluency	.60**	.55**	—
4. Comm. adequacy	.54**	.72***	.75***

Note. *** p < .001, ** p < .01. Comm. adequacy = communicative adequacy

Table 44. *Bivariate Correlations Between Fluency and Communicative Adequacy Measures (Time 2)*

	1	2	3
1. Utterance fluency	—		
2. Fluent syllables per min.	.83***	—	
3. Perceived fluency	.71***	.62**	—
4. Comm. adequacy	.63**	.59**	.70***

Note. *** p < .001, ** p < .01. Comm. adequacy = communicative adequacy

Table 45. *Bivariate Correlations Between Fluency and Communicative Adequacy Measures (Time 3)*

	1	2	3
1. Utterance fluency	—		
2. Fluent syllables per min.	.68***	—	
3. Perceived fluency	.65**	.59**	—
4. Comm. adequacy	.48*	.61**	.75***

Note. *** p < .001, ** p < .01, * p < .05. Comm. adequacy = comm. adequacy

Research Question 2: The Extent of Interactional Competence Development

Research question 2 investigated the extent to which the participants' interactional competence developed during a one-semester course of spoken interaction. In addition to the quantitative analysis, answering this question required detailed qualitative analysis of the interactional features produced during the group discussions. It also entailed comparing the speech produced by the student participants with that produced by a group of highly proficient Japanese speakers of English.

Quantitative Analysis

The aim of the quantitative analysis was to examine changes in the use of the various interactional features over the three measurement points. However, the data were not expected to be normally distributed due to the relatively small number of anticipated moves in each category. In addition, the fact that all measures were contingent on the behavior of other participants means that they were not independent. Therefore, it was decided in advance to use non-parametric statistics with these data. Accordingly, Friedman's ANOVAs were conducted for each of the topic development and support move categories. The results are summarized in Table 46, with mean values at each time point presented alongside p-values and effect sizes—in this case Kendall's W—of the associated ANOVA. Statistically significant results were produced for the following categories: topic extension statements; minimal extensions; and turn/topic management. Results are presented in detail for these measures below.

Regarding the other measures, the change in the mean number of topic extension questions was not statistically significant, despite a marked drop of 2.30 between Time 1 and Time 2 (from 5.20 to 2.90) and a reduction of 2.67 overall. This outcome is perhaps explained by the fact that more than 50.00% of the total values were accounted for by only four participants at Time 1, thus inflating the mean at this point, as illustrated by a large standard deviation (5.07) and wide confidence intervals (2.83-7.58). In addition, none of the support moves (backchannels, agree/disagree markers, confirmations of comprehension, and repairs) produced significant changes, indicating that their use was relatively stable throughout the semester, at least in terms of quantity of output.

Table 46. *Summary of Results for All Interactional Features*

	Time 1	Time 2	Time 3	*p*	*W*
Topic ext. statement	8.45	6.55	5.05	< .01*	.28
Minimal extension	3.50	1.65	0.90	< .01*	.30
Topic ext. question	5.20	2.90	2.53	.15	.10
Turn/topic management	1.65	2.75	3.15	< .01*	.23
Backchannels	10.30	10.55	11.05	.86	< .01
Agree/disagree markers	1.20	1.25	0.95	.72	.02
Confirmations	1.85	1.85	2.10	.98	< .01
Repairs	1.35	0.90	0.95	.31	.06

Note. * Denotes statistically significant result. Ext. = extension.

Topic Extension Statements

The mean number of topic extension statements produced was 8.45 at Time 1, 6.55 at Time 2, and 5.05 at Time 3 (see Table 47). Therefore, there was a decrease of 1.90 between Times 1 and 2, and a further drop of 1.50 between Times 2 and 3. The relatively large standard deviations and wide confidence intervals at Times 1 and 2 (*SD* = 3.62 and 3.24, respectively) indicate substantial variance between individuals, although this had narrowed considerably by Time 3 (*SD* = 2.04).

Although the data were not assumed to be normally distributed, it is nevertheless worth noting the presence of an extreme outlier at Time 2. This participant contributed 17 such statements, as opposed to 11 at Time 1 and 5 at Time 3, which was far above the Time 2 mean of 6.55 and well above the second highest value of 10.00. This data point highlights the unstable nature of group interaction, as each instance generates its own dynamics due to the large number of variables involved. Two further outliers were identified at Time 3, although these values, of 11.00 and 9.00, were much closer to the respective means.

Table 47. *Descriptive Statistics for Topic Extension Statements*

	Time 1	Time 2	Time 3
M	8.45	6.55	5.05
SE	0.81	0.72	0.46
95% CI	[6.76, 10.14]	[5.04, 8.06]	[4.10, 6.00]
SD	3.62	3.24	2.04

Note. N = 20 at each time point.

Friedman's ANOVA indicated that these changes were significant, $\chi^2(2) = 11.08$, $p = .004$, $W = .28$. Conover's post hoc comparisons (Bonferroni adjusted), using a rank-biserial correlation as an effect size, revealed a significant difference between Time 1 and Time 3 only ($p = .001$, $r_{rb} = .82$). There was no significant difference between Time 1 and Time 2 ($p = .14$, $r_{rb} = .44$), or between Time 2 and Time 3 ($p = .09$, $r_{rb} = .44$).

Minimal Extensions

The mean number of minimal extensions produced was 3.50 at Time 1, 1.65 at Time 2, and 0.90 at Time 3 (see Table 48). Therefore, there was a decrease of 1.85 between Times 1 and 2, and a smaller drop of 0.75 between Times 2 and 3. The fact that the standard deviation is almost equal to or greater than the mean in each case indicates a large amount of variance between individuals. Although the standard deviation and confidence intervals narrowed over time (*SD* = 3.05, 1.84, and 1.12, respectively), these data are skewed by the fact that half of the participants produced zero minimal extensions at Time 3.

Table 48. *Descriptive Statistics for Minimal Extensions*

	Time 1	Time 2	Time 3
M	3.50	1.65	0.90
SE	0.68	0.41	0.25
95% CI	[2.07, 4.93]	[.79, 2.51]	[.38, 1.42]
SD	3.05	1.84	1.12

Note. N = 20 at each time point.

There were again some noteworthy outliers. Two participants produced 11 and 9 such moves at Time 1, far in excess of the group mean of 3.50. The participant who produced the most minimal extensions at Time 1 remained the only outlier at Time 2, this time producing 8, in contrast to a group mean of 1.65, with the next highest value being only 3.00. However, at Time 3 this participant produced only 1 minimal extension, close to the group mean of 0.90.

Friedman's ANOVA indicated that the changes over time were significant, χ^2 (2) = 12.03, p = .002, W = .30. Conover's post hoc comparisons (Bonferroni adjusted) revealed a significant difference between Time 1 and Time 2 (p = .01, r_{rb} = .70) but not between Times 2 and 3 (p = 1, r_{rb} = .43). The overall difference between Time 1 and Time 3 was p = .002, r_{rb} = .98.

All Topic Extensions

The above results reveal a similar pattern regarding both minimal extensions and topic extension statements, whereby the participants collectively produced fewer such moves over time. Although coded separately, these two categories are related because minimal extensions are simply short (i.e., sub-clausal) forms of topic extension statements. Therefore, in order to examine these two variables simultaneously, independent of length, it was decided to additionally investigate the results produced by combining them into one category called *all topic extensions*.

The mean number of all topic extensions produced was 11.95 at Time 1, 8.70 at Time 2, and 5.95 at Time 3 (see Table 49). This result showed a reduction of 3.25 extensions between Times 1 and 2, and 2.75 between Times 2 and 3. The relatively large standard deviations and wide confidence intervals at each time point emphasize the substantial amount of variance between individuals, although again it narrowed over time (SD = 5.44, 4.51, and 2.50, respectively).

Table 49. *Descriptive Statistics for All Topic Extensions*

	Time 1	Time 2	Time 3
M	11.95	8.70	5.95
SE	1.22	1.01	0.56
95% CI	[9.40, 14.50]	[6.59, 10.81]	[4.78, 7.12]
SD	5.44	4.51	2.50

Note. N = 20 at each time point.

Times 2 and 3 each produced two outliers that exceeded the group means, with two additional outliers observed below the mean at Time 3. In these latter cases, both participants produced considerably fewer extensions at this time point (3 and 5, respectively) than they had at previous time points (12 and 15 extensions at Time 1; 8 and 20 extensions at Time 2). These individual cases exemplify the reduction in topic extension moves that occurred over the semester.

In this case, the significant changes over time, indicated by Friedman's ANOVA, were χ^2 (2) = 10.96, p = .004, W = .27. Conover's post hoc comparisons (Bonferroni adjusted) revealed a significant difference between Time 1 and Time 3 only (p = .002, r_{rb}

= .94). There was no significant difference between Time 1 and Time 2 (p = .20, r_{rb} = .47), or between Times 2 and 3 (p = .20, r_{rb} = .53).

Turn/Topic Management

The other significant result was regarding turn/topic management moves, which in this case revealed an increase in such contributions over time. The mean number produced was 1.65 at Time 1, 2.75 at Time 2, and 3.15 at Time 3 (see Table 50). There was therefore an increase of 1.00 between Times 1 and 2, with a smaller rise of 0.40 between Times 2 and 3. The fact that the standard deviation is greater than the mean at Time 1 again indicates a large amount of variance between individuals, exemplified by the presence of three outliers at this time point. In this case, however, the variance remained relatively stable and did not narrow over time (SD = 1.76, 1.92, and 1.76, respectively).

Friedman's ANOVA indicated that the mean changes were significant, χ^2 (2) = 9.36, p = .009, W = .23. Conover's post hoc comparisons (Bonferroni adjusted) showed a significant difference between Time 1 and Time 2 (p = .03, r_{rb} = .50) but not between Times 2 and 3 (p = .65, r_{rb} = .27). The overall difference between Time 1 and Time 3 was p = .009, r_{rb} = .65.

Table 50. *Descriptive Statistics for Turn/Topic Management Moves*

	Time 1	Time 2	Time 3
M	1.65	2.75	3.15
SE	0.39	0.43	0.39
95% CI	[.83, 2.47]	[1.85, 3.65]	[2.33, 3.97]
SD	1.76	1.92	1.76

Note. N = 20 at each time point.

Highly Proficient Speakers

For comparison purposes, a group discussion conducted by three highly proficient Japanese speakers of English was recorded and analyzed. Table 51 shows the number of interactional features used by each participant, as well as the group mean, and the mean of the student group at Time 3 (i.e., at the end of the course). Regarding topic development moves, the highly proficient speakers produced a relatively high number of topic extension statements on average. Each member also contributed at least two turn or topic management moves. However, they produced few minimal extensions and there was an almost total absence of topic extension questions. Regarding support moves, the highly proficient speakers produced a relatively large, and equally distributed, number of backchannels. There were also several agree/disagree markers and confirmations of comprehension. On the other hand, zero repair moves were observed during this discussion.

When compared with the student participants at the end of the course, the highly proficient speakers produced more topic extension statements on average (8.33 vs. 5.05). They also produced more backchannels (16.00 vs. 11.05). However, they asked fewer topic extension questions (0.33 vs. 2.53). They also engaged in less negotiation of meaning, both in terms of confirmations of comprehensions (1.00 vs. 2.10) and repairs of misunderstandings (0.00 vs. 0.95). For the remaining categories (i.e., minimal extensions, turn/topic management, and agree/disagree markers), both the highly proficient speakers and the student participants produced broadly similar results. The implications of this comparison are examined in Chapter 6.

Table 51. *Interactional Moves of Highly Proficient Speakers*

	Part. A	Part. B	Part. C	Total	*M*	ST3*M*
Topic ext. statement	11	10	4	25	8.33	5.05
Minimal extension	0	0	2	2	0.67	0.90
Topic ext. question	0	1	0	1	0.33	2.53
Turn/topic management	3	3	2	8	2.67	3.15
Backchannels	16	17	15	48	16.00	11.05
Agree/disagree markers	0	3	1	4	1.33	0.95
Confirmations	2	0	1	3	1.00	2.10
Repairs	0	0	0	0	0.00	0.95

Note. * Part. = participant, ext. = extension. ST3*M* = mean of student group at Time 3

Qualitative Analysis

The qualitative component of the analysis is intended to investigate how effectively and appropriately the various interactional features were used, and whether development could be observed—among individual participants and the group as a whole—during the semester. Pseudonyms have been used, although only the initials of non-core participants are used in the excerpts. The features are presented and analyzed individually, although some categories with overlapping functions have been combined. Of the topic development moves, topic extension statements and minimal extensions have been combined into one category called *topic extensions* because, as mentioned above, the latter are sub-clausal versions of the former and therefore perform a similar function. Regarding the support moves, confirmations of comprehension and repairs have been combined into one category called *negotiation of meaning*, as these features often occur within the same sequences and some forms can be used for both purposes. Backchannels have not been analyzed in this section because there is limited scope for the qualitative development of such tokens as “Yes” and “Okay”; therefore, quantitative analysis is sufficient for the purposes of this study. Further discussion and interpretation of the themes identified in this chapter is presented in Chapter 6.

Topic Extensions

The first interactional feature examined is topic extensions, as they provide the majority of content in a group discussion. The main theme that emerged regarding these features was a general shift from short, sometimes single-word, extensions to fewer but longer contributions that addressed the topics under discussion more fully and incorporated more details and supporting reasons. The following exemplars, based on the contributions of two participants (Airi and Naho), illustrate this pattern.

Beginning with Airi, at Time 1 (i.e., at the beginning of the course) she answered the discussion question *What is the best season in Japan?* with the following topic extension, consisting of an opinion and a single reason:

Excerpt 1 (Time 1)

Airi: I think winter is (.) the best season in Japan? (.6) because I like skiing? (.8) so (1.0) yeah.

Her initial comment was followed up later in the discussion during this exchange, in which she offered one single-clause extension (line 2), two minimal (i.e., sub-clausal) extensions (lines 4 and 8), and one multi-clausal extension (line 6):

Excerpt 2 (Time 1)

1 I: Where do you (.8) go (.6) in winter.

2 Airi: Oh I want (.5) to go (.) Nagano.

3 I: Ah: ski:

4 Airi: Ski- to ski

(7.1)

5 I: Can you (.) skiing?

(1.1)

6 Airi: I started (.5) I started to ski? (.6) when I was (.9) five? (.) years old

(2.6) so I (1.4) went to ski every year

(3.2)

7 I: Can you: sports ano- another sports

(3.0)

8 Airi: Tennis

In contrast, at Time 3 (i.e., at the end of the course) she answered the discussion question *What is the best place to visit in Japan?* with the following extended multi-clausal contribution:

Excerpt 3 (Time 3)

Airi: I think (.) Okinawa is the best place (.) to visit because er (.) I visit (.) er Okinawa twice? first is (.6) school trip? and the other is: family trip (.7) and (2.2) er:: (1.3) in Okinawa there were a lot–there were museum about World War Two? (.5) so I learned a lot of things? (1.22) in the (.) museum and (1.4) and I learned about World War Two.

Later in the same discussion Airi produced another multi-clausal extension, bisected by a backchannel, before finishing with a turn-taking question:

Excerpt 4 (Time 3)

1 Airi: I think (.7) the sea is better (.7) because (.) I (.) don't like (.) insect, insects

2 Kaito: Insects ah:

3 Airi: there (1.2) there are a lot of insect in mountain so (1.26) um (.6) and also (1.1) I learned swimming (.) when I was a little child so (1.1) I like (.) swimming in the sea (1.4) what does everyone think?

Following a similar pattern, at Time 1 Naho was involved in this exchange, in which she provided a series of short topic extensions (lines 1, 7, 11, and 13) when asked about the best place to visit in Japan.

Excerpt 5 (Time 1)

1 Naho: I- I (1.9) I want- mm? I like (.8) Okinawa is the best place to [visit in Japan.

2 Tetsuya: Oh::] Okinawa.

3 Naho: Yes

4 Manabu: I have never been to Okinawa. ((laughs))

5 Naho: [Oh really?

6 Tetsuya: Yeah me too]

7 Naho: Wow (1.0) you should go there, ((laughter)) yes

8 Manabu: You like swimming right?

9 Naho: Er::: ah:: [so so

10 Manabu: Not swimming?] ((laughter))

11 Naho: but I like watch sea- watch the sea beautiful sea

12 Manabu: Beautiful sea

13 Naho: Yes: Okinawa is (.5) the best (.6) photo spot.

At Time 3, however, she produced this extended multi-clausal contribution in answer to a question about high school life:

Excerpt 6 (Time 3)

1 Naho: For me: I had a wonderful experience in my high school so (.6) I really enjoyed (.5) high school life? so one of my beautiful experience of (.5) my high school is (.) er club? (.8) team so (.) I: belonged to (.8) baton dance club? so: we do: (1.0) baton and dance- both baton and dance so (.9) I often belonged to dance team so (.6) er:: we often (.) practice dance together and then (.6) eat (.) dinner or lunch or play?

2 M: [Wow

3 Naho: so] that was interesting for me? and I was able to make a lot of friends there? (.6) so that was so: (1.1) so wonderful memory?

4 Konomi: Yes

5 Naho: for me I think?

The above excerpts focus on two participants to exemplify the general development of the group as a whole, which saw a shift from serial, short topic extensions in the early part of the course to fewer but more detailed, multi-clausal contributions by the end.

Topic Extension Questions

In keeping with the theme of fewer topic extensions over time, fewer topic extension questions were also observed at Times 2 and 3 (i.e., the middle and end of the course) than at Time 1. The disparity in this case was mainly due to the large number of such questions asked by a small number of participants at Time 1. The following excerpts involve two of those participants, Rie and Akari, and illustrate the relatively superficial nature of many of the discussions at this time point.

In Excerpt 7, Rie asked three questions in succession (lines 1, 3, and 6), the last of which involved a change of topic following an extended pause.

Excerpt 7 (Time 1)

1 Rie: Is <u>there</u> cherry blossom in Korea?

2 S: Yes

3 Rie: Ah yes (1.0) many?

4 S: Um:: many but (.5) smaller than.

5 Naoya: Beautiful

(13.4)

6 Rie: Mm ah (.) what food do you like in summer?

Excerpt 8 shows Akari also asking three questions in succession (lines 1, 4, and 6). They are all related to the same topic, although the first two are delivered after two false starts. The fact that each one is answered by a minimal extension (lines 2, 5, and 7) emphasizes the perfunctory nature of many discussions in the early part of the course.

Excerpt 8 (Time 1)

1 Akari: Di- ha- (.) have you ever been to some festival?

2 Kiyomi: Yes

3 L: Ah yes

4 Akari: Wha- w- (.) where::

5 Kiyomi: Where, ah::: maybe Yokohama.

6 Akari: Ah:: big festival?

7 Kiyomi: Yes yes yes.

Excerpts 7 and 8 also highlight the contingent nature of interaction, as receiving such short responses prompted Rie and Akari to ask further follow-up questions, which perhaps they would not have felt necessary had they received longer answers. However, both Rie and Akari asked far fewer topic extension questions at Times 2 and 3. Specifically, Rie went from asking 16 at Time 1 to only two at Times 2 and 3, whereas Akari went from 14 at Time 1, to three at Time 2, and nine at Time 3. Furthermore, similar to the trend towards longer topic extension statements that occurred over time, there was also a qualitative shift in the nature of topic extension questions. Although at Time 1 many questions consisted of a single lexical item (e.g., Where?, Why?), questions at Time 3 tended to be more precisely targeted. Examining the development of Aoi's questions illustrates this point.

At Time 1, Aoi asked ten topic extension questions, four of which were one-word *Why?* questions, as in the following example (line 4).

Excerpt 9 (Time 1)

1 M: My best place to visit in Japan is Hakone

2 I: Hakone

3 M: Hakone yes ah::

4 Aoi: Why

At Time 3, however, Aoi asked only four topic extension questions in total, none of which consisted of just a single lexical item. The following sequence contains three of those questions (lines 1, 5, and 9), all based on the same topic, which was thus explored in depth.

Excerpt 10 (Time 3)

1 Aoi: So why change your (.6) <u>club</u>

(3.1)

2 M: I like the cooking club and the English club

3	Aoi:	Yeah
4	M:	but (1.1) when I (.9) when I saw the perform (.) of my dance club? (.6) so cool…
5	Aoi:	Did you have any experience of dancing before you (.5) entered the dancing club
6	M:	No I- I (.6) the first time
7	Manabu:	First time
8	M:	in er high school (.) so I was so nervous …
9	Aoi:	In present situation do you enter the any dance circle in this university?
10	M:	I: mm I try to the (1.7) belong to the %unclear% but (1.1) I don't accept the email.

The above excerpts again focus on individuals to exemplify the general development of the group. They show that topic extension questions followed a similar pattern to topic extension statements in that initially sub-clausal or lexical items were replaced by more precise and detailed questions towards the end of the course.

Turn/Topic Management

In addition to an increase in the number of turn and topic management features used over time (see Table 50 above), the main qualitative theme that emerged was a gradual progression towards the smoother use of such features and a reduced reliance on gesture or L1 Japanese. At Time 1, the participants often relied on non-verbal means when transitioning between turns and topics, such as silently pointing to themselves or raising a hand to indicate the desire to take a turn and gesturing the number two to suggest moving onto the second discussion question. In other cases, Japanese (i.e., the dominant language of these participants), was used to achieve these purposes, sometimes accompanied by nervous laughter. Often the Japanese was whispered or mumbled at low volume, implying that the participants felt they were transgressing classroom norms by doing so. In one instance the Japanese question *tsugi iku?*, which translates approximately as *Shall we move on to the next one?* was audible. Moreover, when turn taking was attempted in English at this time point it was often done so clumsily, as in the following excerpt:

Excerpt 11 (Time 1)

1 Katsuya: Er::m can I next? (1.0) erm
((Konomi gestures to S))

2 Katsuya: I think- ah ((Katsuya gestures to himself)) o- can I?
((S gestures to Katsuya))

3 Konomi: Yeah yeah ((gestures to Katsuya))

4 Katsuya: okay? (.) er: my t- (.5) my best favorite place is (.6) I think
Okinawa…

At Time 2 (i.e., in the middle of the course), there was much less reliance on Japanese to signal turn taking or changing topic. Nevertheless, a reliance on gesture was still evident at times, and the English used to manage these transitions often included disfluent features, such as pauses (see Excerpt 12), self-repairs (Excerpt 13), or false starts (Excerpt 14).

Excerpt 12 (Time 2)

Tetsuya: Which (.) do you think (.6) is better summer or winter? (2.2) something ((gestures to other participants))

Excerpt 13 (Time 2)

1 I: What is the best season in Japan

(1.1)

2 Konomi: Ah:: can I say start? ah can I start (.9) ah::: I think? (.5) the best season in Japan is fall

Excerpt 14 (Time 2)

Akari: Which do you: think is better- ah the second question is er which do you think is better high school or university <u>life</u> (1.8) what does everyone think? (3.8) can I (.5) start?

Although not always produced fluently, Excerpts 13 and 14 also contain the formulaic phrases *Can I start?* and *What does everyone think?* to signal either the desire to take a turn or to invite others to do so. However, by the end of the course, turn taking and topic management had generally become smoother, such that it was common to hear these formulaic phrases used appropriately and relatively fluently, as demonstrated in the following excerpt (lines 2 and 3), in which all three members collaborated to begin the discussion:

Excerpt 15 (Time 3)

1 Shinpei: What is the best place to visit in Japan.

(2.8)

2 Akari: What does everyone <u>think</u>

(1.8)

3 C: Can I start?

4 Shinpei: [Sure

5 Akari: <u>Sure</u>]

6 C: Er:: I think the best place to visit in Japan is Kyoto?

7 Akari: Ah yeah

8 C: because there (.7) is a lot of (.7) his- beautiful historical buildings?...

Moreover, the following Time 3 excerpt contrasts with the Time 1 instance mentioned earlier in which the transition between topics was conducted solely in Japanese (i.e., *tsugi iku?*)

Excerpt 16 (Time 3)

1 Aoi: So (.) we move on to next question?

2 N: [Ok

3 Manabu: Yeah]

4 Aoi: Which do you think is better high school or university life.

Finally, as with the previous interactional features, examining the same individual at the beginning and end of the course provides a clear illustration of development. In the following Time 1 excerpt, although Midori took the initiative in introducing a new topic, she did so by looking down at the question paper and asking the truncated question "Summer or winter?" to the group, then later initiated the following perfunctory exchange about favorite seasons:

Excerpt 17 (Time 1)

1 Midori: ((gestures to K)) Spring?

2 Kaito: Spring er win- winter [winter

3 Midori: Winter] winter

4 C: ((looks at M)) *are?* (huh?) Summer

5 Midori: Summer

6 C: Summer ((laughter))

In contrast, in the following Time 3 excerpt Midori introduced a new topic by reading the prepared discussion question in full and following up with a turn-taking invitation, leading to a more expansive exchange:

Excerpt 18 (Time 3)

1 Midori: What is the best place to visit in Japan (.6) what does everyone think

(2.8)

2 Kenji: Can I say something?

3 Midori: [Yes

4 Rie: Sure]

5 Kenji: Oh:: (.5) I think er:: Kyoto is the best place to visit in Japan (.5) oh: because er there are a lot of (.) er historical buildings?...

The above excerpts illustrate the development of turn taking and topic management. Specifically, by the end of the course there was a reduced reliance on L1 Japanese and non-verbal means during such transitions, and an increased use of formulaic phrases.

Agree/Disagree Markers

The main theme that emerged regarding the use of agree/disagree markers was a gradual diversification (i.e., increased range) of methods used over time. At Time 1, of the 24 such markers used by the 20 core participants, 12 were simple *me too* tokens. For example, all four of the agreement markers Kiyomi used at Time 1 were of this type, as in the following excerpt:

Excerpt 19 (Time 1)

1 R: I think (.5) spring (.) is the best.

2 Kiyomi: Me too.

Furthermore, Kiyomi did not expand upon any of these markers unless prompted to do so. At Time 2, however, she offered a more expansive variation in which the agreement marker *too* was embedded within a topic extension statement:

Excerpt 20 (Time 2)

1 Kiyomi: I think (.7) university life is better than high school too

2 Shinpei: Yeah

3 Kiyomi: er: because (1.1) in my experience mm (1.1) mm we can (1.0) behave very <u>free</u> (.6) freely…

The second most common agreement marker at Time 1 was the formulaic phrase *I think so too*, which was used six times. However, on three of those occurrences it was produced without the requisite adverb *too*, as in the following excerpt:

Excerpt 21 (Time 1)

1 L: …but winter is more great.

2 Akari: Ah:: I think so.

By the second recording point, *I think so too* had overtaken *me too* as the dominant method of expressing agreement. Specifically, *I think so too* accounted for 11 of the 25 agreement markers at this point, whereas *me too* was used only three times. Furthermore, unlike at Time 1, it was produced accurately on each occasion, as in this instance in which it was followed up by a single-clause topic extension:

Excerpt 22 (Time 2)

1 Katsuya: …so summer is (.) better I think (.5) yeah

2 Naho: Mm (1.3) yeah: I think so too (1.2) winter is so <u>cold</u>

This diversification of techniques also included a progression from relatively abrupt expressions of disagreement to more indirect or nuanced forms. For example, at Time 1 Tetsuya was involved in the following exchange:

Excerpt 23 (Time 1)

1 Manabu: <u>Moun</u>tain is dangerous.

2 Tetsuya: <u>No</u> ((shakes head))

3 Manabu: No dangerous?

4 Tetsuya: <u>No</u> dangerous <u>no</u> dangerous…

In contrast, at Time 2 Tetsuya attempted a more indirect method, although lacking in grammatical accuracy, in which he attempted to mitigate his disagreement (line 4) while explaining his positive assessment of summer:

Excerpt 24 (Time 2)

1 C: …I like (.8) only I like autumn's temperature.

2 Tetsuya: Oh:: okay okay [okay

3 C: Yeah]

4 Tetsuya: I understand that you said but (1.6) summer's (.) insects? and summer's temperature

5 C: Yeah

6 Tetsuya: is a <u>spice</u>: ((laughter))

7 C: [A spice

8 Tetsuya: spice] for to enjoy the summer.

At Time 3, a more accurate attempt at mitigated disagreement, albeit repeating the word *opinion*, was produced by Midori, who had not used any agree or disagree markers in the previous two recordings, at the beginning and middle of the course. She then used this marker to lead into a multi-clausal topic extension statement.

Excerpt 25 (Time 3)

Midori: I understand your opinion (.7) opinion (.5) but (.6) er:: I prefer the mountains better than visiting the sea? (1.0) er because in the mountain er we can enjoy the fresh air in the mountain? (1.4) a:nd er (.8) mm: the mountain has a (.) cold- cold? cold temperature than (.5) sea?...

By Time 3, diversification had progressed to the extent that there was no dominant form of indicating agreement or disagreement, in contrast to Times 1 and 2. Of the 19 such markers used by the core participants at Time 3, no single form was attempted more than four times. This progression is exemplified by Kaito, whose use of such markers gradually became more nuanced and diversified over the three recording points. At Time 1, the only marker he produced was the short form *me too*:

Excerpt 26 (Time 1)

1 C: I think the best season in Japan is (.9) er spring

2 Kaito: Ah me too

At Time 2, he also produced a single expression of agreement, although on this occasion it was embedded within a multi-clausal extension:

Excerpt 27 (Time 2)

Kaito: I think (2.3) er:: similar to Himari? (1.1) both (.7) high school life and university life is (.) good for students? er but I think university life is better than high school (.) life (.9) because er we can er many (.9) many free time (.5) in university…

Finally, at Time 3 he produced three such markers, one of which was similar to the Time 2 form, beginning "Similar to Risa, I recommend Kyoto." A second consisted of the short response "That's right." However, the third was a nuanced indication of disagreement, mitigated by partial agreement with the previous speaker (line 5), preceded by a turn-taking request (line 2), and followed by a multi-clausal extension (line 7).

Excerpt 28 (Time 3)

1 L: …I- I think the sea: (.7) is (2.1) mm for me is good.

(1.6)

2 Kaito: I see (1.2) er I er can I say something?

3 L: [Sure

4 Airi: Sure]

5 Kaito: Of course I like sea and [swimming

6 Airi: Okay] ((laughter))

7 Kaito: but (.5) I- I recommend mountain? er mountains er because er I hate insects but er I like camp (1.3) camp in mountains (.7) I recommend er (.9) doing camp in mountains and er you er (2.0) can refresh? refresh mind in (.8) er clean air in mountains and (2.5) by doing walking? er it is (.5) good exercise for (.6) health…

The above excerpts illustrate the development of agree/disagree markers and the gradual diversification of such techniques, including more nuanced and mitigated forms of disagreement, that had taken place by the end of the course.

Negotiation of Meaning

The primary themes that emerged regarding negotiation of meaning—which combines the categories of confirmations and repairs—were a reduced reliance on

Japanese and the application of negotiation techniques to a broader range of situations. However, due to the highly collaborative and contingent nature of such sequences, it is more difficult to track individual progress regarding these features than those analyzed previously. Not only is their occurrence inherently less predictable than the need to generate or respond to a topic extension, for example, but participants must perceive a trouble source as problematic before it can be negotiated (Wong & Waring, 2021). Nevertheless, changes in the way specific confirmation or repair features were used over time reveal general trends within the group.

First, the lack of significant change regarding quantification over time (see Table 46) implies that some participants were negotiating meaning in English from the earliest stages of the course. In particular, there were numerous instances at Time 1 of students offering vocabulary items following a breakdown, as well as confirming the meaning of a Japanese item in English. For example, in the following sequence, Midori provided five vocabulary items in quick succession (lines 2, 4, 10, 14, and 16) following breakdowns or pauses, all of which were taken up by the speaker, thus enabling the discussion to progress.

Excerpt 29 (Time 1)

1	C:	And er (.8) personally I wan- I:: like (.) winter (.) clothes, (.8) winter clothes is er (1.1)
2	Midori:	Cute?
3	C:	cute ((laughter)) cute
		(1.1)

4 Midori: Cute and cool

5 C: Cool ((laughter)) better than summer

(7.1)

6 Midori: <u>And</u>

7 C: And

8 Midori: <u>and</u> (.6) er::

(11.3)

9 C: %whispers inaudibly in Japanese% spring summer *aki aki aki*

(fall fall fall)

10 Midori: Fall

11 C: Fall

12 Midori: In fall?

13 C: In fall (.5) in fall (1.5) weather is er

14 Midori: Comfortable

15 C: comfortable (1.7) comfortable (.9) and er (1.6) *nan daro* (3.5)

undokai nanto iu no (what is it sports day how do you say that)

exer-

16 Midori: Sports sports [<u>fes</u>tival.

17 C: ah sports] festival

On the other hand, some participants contributed few negotiation moves throughout the course. For example, Kenji provided a total of two confirmations and no repairs, while Ryota offered one repair and no confirmations during the three recordings.

Nevertheless, when participants did engage in meaning negotiation a reduced reliance on L1 Japanese was evident over time. At Time 1, using Japanese was often the initial method of dealing with a trouble source, with participants then collaboratively attempting to produce an English equivalent, as shown in lines 9 and 15 in Excerpt 29 (above). Moreover, in some instances Japanese use was not treated as problematic and no negotiation even took place. The following excerpt provides one such example, in which no attempt was made to either elicit or confirm the English meaning of a Japanese vocabulary item:

Excerpt 30 (Time 1)

1 S: But in the summer (.) er bad point (.) is er (1.0) *nechusho* (heat stroke)

2 Honoka: Ah *nechusho*

Japanese was less prevalent in general at Time 2 (i.e., the mid-point of the course) and tended to be used less overtly than at Time 1. Furthermore, in contrast to the above excerpt, it was almost always treated as a trouble source that needed to be negotiated, as in the following sequence:

Excerpt 31 (Time 2)

1 Honoka: Did- did you:: go:: (.7) *eto nan da ke* (er what is it) t- (.9) travel *janakute shugakuryoko* (I mean school trip)

2 Kenji: Ah:: travel- travel?

3 Honoka: Travel

4 Kenji: School travel?

5 Honoka: School travel school travel

6 Kenji: School trip school trip

7 S: Ah school trip

8 Honoka: School trip school trip ((laughter))

By Time 3 there was little Japanese use during the discussions, even after a trouble source had been identified. When necessary, the participants tended to rely on gesture rather than L1 to supplement the target language and convey meaning during negotiation sequences, as in the following excerpt:

Excerpt 32 (Time 3)

1 Y: Girls er played ah volleyball er but er I (.) dislike volleyball?

2 M: Why

3 Y: er (1.0) hand is er (.5) hand is ((gestures hitting hand))

4 Tetsuya: Ha- your hand [hurt?

5 Y: sore] sore sore sore er hand is sore sore?...

Apart from a gradual avoidance of Japanese when dealing with communication problems, the other theme identified in this section was the application of negotiation techniques to a broader range of situations. The technique of rephrasing a prior comment, and how it was used to negotiate or clarify meaning at each time point, is used to illustrate this form of development. Rephrasing here is used to mean changing a completed instance of the target language into an alternative form of the target language, either for confirmation or repair purposes. It can involve rephrasing either one's own or another participant's utterance. However, it does not include instances of converting L1 into L2, such as in Excerpt 31, or offering vocabulary items after an incomplete word search, as in Excerpt 29.

The only instances of such rephrasing at Time 1 involved substituting a single item in order to confirm the meaning of a previous utterance. For example, in the following excerpt Rie replaced her initial word choice *relative* with *cousin* at line 5 after it became clear that S had not understood the question.

Excerpt 33 (Time 1)

1	Rie:	Many relative?
2	S:	Hm?
3	Rie:	Relative? (1.1) ((turns to N)) relative
4	Naoya:	Relative ((nods))
		(1.9)
5	Rie:	((turns back to S)) Cousin?
6	S:	Ah no

There were also several similar instances at Time 2, in which a single item was substituted for confirmation or repair purposes. However, there was also a more elaborate occurrence at this time point, in which an interlocutor (Aoi) rephrased her own question over several turns (Excerpt 35, lines 1, 5, 7, and 9) due to the apparent non-comprehension of the recipient.

Excerpt 34 (Time 2)

1 Aoi: …so do you have any hobbies in (.7) fall?

2 Y: Fall?

3 Aoi: Yeah

4 Y: Fall?

5 Aoi: Yeah did you ha- did you have

6 Y: Ah yes yes

7 Aoi: did you get some new hobby in fall (.5) last year?

8 Y: Ah get new hobby?

9 Aoi: Yeah or try something like that.

This technique was extended further at Time 3, as there were two occurrences that served a different purpose from those previously mentioned. Whereas at Times 1 and 2 rephrasing was used to either clarify meaning or repair a communication breakdown, these two novel instances involved summarizing and reformulating a previous speaker's opinion, either as a precursor to changing the topic (Excerpt 35, line 4) or to offering a contrasting opinion (Excerpt 36, line 2).

Excerpt 35 (Time 3)

1 Emiri: … all season we (.) played the (.8) guitar.

2 Keita: Yeah

3 Emiri: Yeah it was enjoyable.

4 Keita: So you mean- do you mean your high school environment is special and (1.0) and several activities makes you happy?

5 Emiri: <u>Yeah</u> (.5) that's (.) [right

6 H: Nice]

7 Keita: Okay?

8 Emiri: Okay

9 Keita: Next topic?

10 Emiri: Yes

Excerpt 36 (Time 3)

1 Naho: … my teacher always give me interesting story and we often do experiments?... yes that makes me so [happy.

2 Konomi: You] (.6) you means you can en<u>joy</u> (.7) your high school life?

3 Naho: <u>Yes</u> (.7) how about <u>you</u>.

(1.3)

4 Konomi: Yeah I: (1.3) I:'m not (1.8) I'm not enjo:y I cannot enjoy my high school life?...

In addition, the range of objects of repair expanded over time. At Time 1 they consisted almost exclusively of individual lexical items (see Excerpt 29 above), but by Time 3 they included the collaborative completion of whole clauses. In total, there were four instances at Time 3 of other-initiated repair in which the listener offered more than just a single lexical item. All four are shown below—the repair initiation comes at line 2 in Excerpts 37 and 38, and line 4 in Excerpts 39 and 40. Each repair initiation followed a noticeable pause, indicating a breakdown of communication, and was taken up by the speaker in each case (albeit only partially in Excerpt 38).

Excerpt 37 (Time 3)

1 H: After school we:: (1.9) play table tennis? or: (.9) *karuta*? (cards) (.5) or: (.) basketball so we: (2.3) ((laughs nervously)) we::

2 Keita: Enjoy [the match?

3 H: Yes]

Excerpt 38 (Time 3)

1 L: I: (.7) agree with you. I also hate the insects. ((laughter)) okay because my- my hometown is nearby a mountain? so there are (1.6)

2 Kaito: lot of [insects?

3 L: lot of] mosquitoes.

Excerpt 39 (Time 3)

1 C: Maybe the (2.4) sho- (.5) shop er (1.4) *nani* (what) older (.6) person-

2 Katsuya: ((quietly)) Older

3 C: older person (1.4) i:s (1.3) so (1.9)

4 Katsuya: Not good [person?

5 C: Yeah] not good

Excerpt 40 (Time 3)

1 Y: And er when I was first grade high school student?

2 Tetsuya: Yes

3 Y: er there is a chorus er concert (1.6) er but er second and third grade is er (.5) it is er not h- (1.3) it is not [hold

4 Tetsuya: Cannot] take place

5 Y: ca- cannot take

The above excerpts illustrate the development of meaning negotiation throughout the course, including a reduced reliance on Japanese when dealing with trouble sources. The example of rephrasing was used to show the gradual application of negotiation techniques to a broader range of situations, from substituting single lexical items at the beginning of the course to producing summative reformulations by the end. Collaborative completions were also used to illustrate the expanded range of meaning negotiation techniques.

Research Question 3: The Extent of Speaking Self-Efficacy and Anxiety Change

Research question 3 investigated the extent to which the participants' speaking self-efficacy and anxiety changed during a one-semester course of spoken interaction. One-way repeated measures ANOVAs were conducted to examine the extent of any such changes. Both the self-efficacy and anxiety measures were based on Rasch analyses of responses to Likert-scale questionnaires administered at each data collection point. Validation evidence for all individual difference measures was presented in Chapter 4. The results are summarized in Table 52, with mean values at each time point presented alongside *p*-values and effect sizes of the related ANOVA. The results for each variable are then presented in detail.

Table 52. *Summary of Speaking Self-Efficacy and Anxiety Results*

	Time 1	Time 2	Time 3	p	η^2
Self-efficacy	1.29	2.43	2.70	< .01	.27
Anxiety	0.56	0.16	-0.07	.21	.08

Note. * N = 20 at each time point. All statistics are based on Rasch logits.

Speaking Self-Efficacy

Looking at speaking self-efficacy first, changes in the mean values (see Table 52) show an increase of 1.14 logits between Times 1 and 2 but only 0.27 logits between Times 2 and 3. The Shapiro-Wilk test confirmed that the data were normally distributed at each time point (p = .64, .76, and .21, respectively). There were no outliers at any time point. Refer to Table 18 in Chapter 4 for the descriptive statistics. Mauchly's test indicated that the assumption of sphericity was met, with non-significant results for the Chi-square test $\chi^2(2) = 0.30$, $p = .86$.

The repeated-measures ANOVA revealed statistically significant changes in speaking self-efficacy over time, $F(2, 38) = 7.07$, $p = .002$, $\eta^2 = .27$, with 27.00% of the variance accounted for by the model. Post hoc *t*-tests (Bonferroni adjusted) revealed a significant pairwise difference between Time 1 and Time 3 ($p = .003$, $d = 0.68$), and Time 1 and Time 2 ($p = .02$, $d = 0.55$), indicating a small effect size in each case, but not between Times 2 Time 3 (p = 1, d = 0.13). Therefore, the participants' speaking self-efficacy increased during the course, specifically during the first half of the semester.

Speaking Anxiety

Looking at speaking anxiety next, the Shapiro-Wilk test confirmed that the data were normally distributed at each time point (p = .34, .82, and .48, respectively). There was one outlier at Time 2 only, with this participant reporting much lower anxiety at this point than at Times 1 and 3. However, as they did not answer all questions identically there was no reason to suppose that these scores were given without due consideration; therefore, it was decided to retain them in the analysis. Refer to Table 22 in Chapter 4 for the descriptive statistics. Mauchly's test indicated that the assumption of sphericity was met, with non-significant results for the Chi-square test $\chi^2(2) = 1.13$, $p = .57$.

The repeated measures ANOVA revealed no significant changes in speaking anxiety over time, $F(2, 38) = 1.65$, $p = .21$, $\eta^2 = .08$, with only 8.00% of the variance accounted for by the model. Post hoc t-tests (Bonferroni adjusted) confirmed no significant change between Time 1 and Time 3 (d = 0.37). Although anxiety fell slightly more between Times 1 and 2 (d = 0.23) than between Times 2 and Time 3 (d = 0.14), neither of these differences reached statistical significance. It can therefore be concluded that, whereas the participants' speaking self-efficacy increased during the course, speaking anxiety remained relatively stable throughout.

Research Question 4: The Extent to Which Speaking Fluency Is Related to Interactional Competence

Research question 4 investigated the extent, if any, to which speaking fluency was related to interactional competence. To answer this question, a bivariate correlation analysis was conducted using the three individual measures of fluency—utterance, perceived, and syllables per minute—as well as the quantitative measures of interactional competence. Regarding the fluency measures, utterance fluency was based on the monologue activities, whereas perceived fluency and syllables per minute were based on the group discussions. Regarding the quantitative measures of interactional competence, prior to analysis the eight categories were aggregated into the two domains of topic development and support moves, in order to reduce the risk of a Type I error. In total, five variables were correlated at each time point.

Tables 53 to 55 show the results, although a separate analysis including the three fluency variables was reported earlier and shown in Tables 43-45. Correlations between these variables are therefore not commented on in this or subsequent sections of this chapter. Regarding interactional competence, the use of topic development and support moves were significantly correlated at each time point. The correlation coefficients are $r = .62$ at Time 1, $r = .71$ at Time 2, and $r = .50$ at Time 3, indicating a large effect at the first two time points and a medium effect at the third.

Table 53. *Bivariate Correlations Between Fluency and Interactional Competence Measures (Time 1)*

	1	2	3	4	5
1. Utterance fluency	—				
2. Fluent syllables per min.	.76***	—			
3. Perceived fluency	.60**	.55*	—		
4. Topic development	.21	.47*	.12	—	
5. Support moves	.35	.60**	.22	.62**	—

Note. $N = 20$. *** $p < .001$, ** $p < .01$. * $p < .05$

Table 54. *Bivariate Correlations Between Fluency and Interactional Competence Measures (Time 2)*

	1	2	3	4	5
1. Utterance fluency	—				
2. Fluent syllables per min.	.83***	—			
3. Perceived fluency	.71***	.62**	—		
4. Topic development	.69***	.63**	.58**	—	
5. Support moves	.56*	.58**	.43	.71***	—

Note. N = 20. *** p < .001, ** p < .01. * p < .05

Table 55. *Bivariate Correlations Between Fluency and Interactional Competence Measures (Time 3)*

	1	2	3	4	5
1. Utterance fluency	—				
2. Fluent syllables per min.	.68***	—			
3. Perceived fluency	.65**	.59**	—		
4. Topic development	.16	.39	-.03	—	
5. Support moves	.17	.22	.05	.50*	—

Note. N = 20. *** p < .001, ** p < .01. * p < .05

Regarding the relationship between fluency and interactional competence, which is the focus of this research question, results are mixed based on these quantitative measures. Significant correlations were found between the use of topic development moves and fluent syllables per minute at Times 1 and 2 (r = .47 and .63, respectively, indicating medium and large effect sizes), although not at Time 3 (p = .09) despite a medium effect size of r = .39. A similar pattern was found regarding the relationship between support moves and syllables per minute, with significant correlations, and large effect sizes, at Times 1 and 2 (r = .60 and .58), although not at Time 3 (r = .22). These results indicate a positive relationship between the number of fluent syllables that the participants produced during discussions and the number of moves they contributed.

However, there is limited evidence to suggest a relationship between quantitative measures of interactional competence and the other aspects of fluency measured here.

Regarding utterance fluency, the trend was towards a positive relationship although significant correlations existed only at Time 2 (r = .69 for topic development moves and r = .56 for support moves, indicating large effect sizes). Regarding perceived fluency, a large and significant positive effect was observed with topic development moves (r = .58) and a medium, although non-significant, positive effect was observed with support moves (r = .43), with inconsistent patterns at the other time points.

Research Question 5: The Extent to Which Speaking Fluency Is Related to Speaking Self-Efficacy, Speaking Anxiety, and Extraversion

Research question 5 investigated the extent to which speaking fluency is related to the individual difference variables of speaking self-efficacy, speaking anxiety, and extraversion. Bivariate correlation analysis was again conducted, using the three individual measures of fluency mentioned previously and the logit measures of the individual difference variables produced by Rasch analysis. Although the extraversion questionnaire was only distributed at Time 1, the results were included in the analyses at all three time points, as dispositional personality traits are considered more stable than contextualized characteristics such as self-efficacy and anxiety. In total, six variables were correlated at each point to answer this research question.

Looking at the relationships among the individual difference variables first, Tables 56 to 58 show a weak to moderate negative relationship between speaking self-efficacy and anxiety (r = -.16., -.47, and -.34, respectively), which reached statistical significance at Time 2 only. A negative, although non-significant, relationship was also found between speaking anxiety and extraversion at Times 1 and 2 (r = -.31. and -.32 respectively, indicating small effect sizes), although not at Time 3 (r = .08). In contrast, a small to medium-sized, positive relationship was found between extraversion and speaking self-efficacy (r = .30, .52, and .25, respectively), although significance was reached only at the second time point.

Table 56. *Bivariate Correlations Between Fluency and Individual Difference Measures (Time 1)*

	1	2	3	4	5
1. Utterance fluency	—				
2. Fluent syllables per min.	.76***	—			
3. Perceived fluency	.60**	.55*	—		
4. Self-efficacy	-.04	.15	.15	—	
5. Anxiety	-.24	-.33	-.07	-.16	—
6. Extraversion	.32	.58**	.33	.30	-.31

Note. N = 20. *** p < .001, ** p < .01. * p < .05

Table 57. *Bivariate Correlations Between Fluency and Individual Difference Measures (Time 2)*

	1	2	3	4	5
1. Utterance fluency	—				
2. Fluent syllables per min.	.83***	—			
3. Perceived fluency	.71***	.62**	—		
4. Self-efficacy	.24	.46*	-.05	—	
5. Anxiety	-.49*	-.66**	-.13	-.47*	—
6. Extraversion	.57**	.68***	.29	.52*	-.32

Note. N = 20. *** p < .001, ** p < .01. * p < .05

Table 58. *Bivariate Correlations Between Fluency and Individual Difference Measures (Time 3)*

	1	2	3	4	5
1. Utterance fluency	—				
2. Fluent syllables per min.	.68***	—			
3. Perceived fluency	.65**	.59**	—		
4. Self-efficacy	.04	-.07	.20	—	
5. Anxiety	-.02	-.07	-.01	-.34	—
6. Extraversion	.44	.63**	.36	.25	.08

Note. N = 20. *** p < .001, ** p < .01. * p < .05

Concerning the relationship between the individual difference variables and fluency, which is the focus of this research question, the clearest association was found between extraversion and fluent syllables produced per minute in the group discussions. In this case, a significant and large positive relationship was found at each time point (r = .58., .68, and .63, respectively). A small to medium positive relationship was also found between extraversion and utterance fluency at each time point (r = .32., .57, and .44, respectively). This relationship only reached statistical significance at Time 2, although a p value of .055 at Time 3 indicates a marginal result likely due to the small sample size. A slightly weaker positive relationship was observed between extraversion and perceived fluency (r = .33. r = .29, and r = .36, respectively), albeit without reaching significance.

Regarding the other two individual difference variables, inverse patterns emerged with respect to their relationships with the fluency variables. Speaking self-efficacy had a positive and significant relationship with fluent syllables produced in the discussions at Time 2 only (r = .46, indicating a medium effect size), while a small, non-significant positive effect was observed with utterance fluency at the same time point (r = .24). In contrast, a negative relationship was found between speaking anxiety and fluent syllables produced at Times 1 and 2 (r = -.33 and -.66, respectively, indicating small and large effect sizes). Again, a smaller effect was observed with utterance fluency at the same time points (r = -.24 and r = -.49, respectively, indicating small and medium effect sizes). However, both relationships were significant only at Time 2. No clear associations were found between either individual difference variable and perceived fluency at any time point.

Research Question 6: The Extent to Which Interactional Competence Is Related to Speaking Self-Efficacy, Speaking Anxiety, and Extraversion

Research question 6 investigated the extent to which interactional competence is related to speaking self-efficacy, speaking anxiety, and extraversion. Bivariate correlation analysis was conducted using the quantitative measures of interactional competence, aggregated into topic development and support moves, and the Rasch logit measures of the three individual difference variables, with extraversion again included at all three time points. A total of five variables were thus correlated at each point to answer this research question. Tables 59 to 61 show the results. The relationships between the two interactional competence variables and the relationships among the three individual difference variables have been reported previously; therefore, correlations between these variables are not commented on in this section.

Concerning the relationship between the individual difference variables and interactional competence, the clearest and most consistent association emerged with respect to extraversion. In sum, a small to medium positive relationship was found between extraversion and both aspects of interactional competence at all three time points. Regarding topic development moves, the correlation coefficients were $r = .36$ at Time 1, $r = .55$ at Time 2, and $r = .44$ at Time 3. Similarly, regarding support moves the coefficients were $r = .38$ at Time 1, $r = .58$ at Time 2, and $r = .30$ at Time 3. Although statistical significance was reached only at Time 2 in both cases, a p value of .051 between extraversion and topic development moves at Time 3 indicate another marginal result probably due to the small sample size.

Table 59. *Bivariate Correlations Between Interactional Competence and Individual Difference Measures (Time 1)*

	1	2	3	4	5
1. Topic development	—				
2. Support moves	.62**	—			
3. Self-efficacy	.25	.40	—		
4. Anxiety	-.16	.05	-.16	—	
5. Extraversion	.36	.38	.30	-.31	—

Note. N = 20. *** $p < .001$, ** $p < .01$. * $p < .05$

Table 60. *Bivariate Correlations Between Interactional Competence and Individual Difference Measures (Time 2)*

	1	2	3	4	5
1. Topic development	—				
2. Support moves	.71***	—			
3. Self-efficacy	.09	.24	—		
4. Anxiety	-.32	-.41	-.47*	—	
5. Extraversion	.55*	.58**	.52*	-.32	—

Note. N = 20. *** $p < .001$, ** $p < .01$. * $p < .05$

Table 61. *Bivariate Correlations Between Interactional Competence and Individual Difference Measures (Time 3)*

	1	2	3	4	5
1. Topic development	—				
2. Support moves	.50*	—			
3. Self-efficacy	-.02	.31	—		
4. Anxiety	-.23	-.46*	-.34	—	
5. Extraversion	.44	.30	.25	.08	—

Note. N = 20. *** $p < .001$, ** $p < .01$. * $p < .05$

Looking at the other individual difference variables, a positive but non-significant relationship was found between speaking self-efficacy and both topic development moves (r =.25) and support moves (r = .40) at Time 1, with a small effect size in the former case and a medium effect size in the latter. Regarding topic development moves, no such relationship was observed thereafter (r = .09 at Time 2 and r = -.02 at Time 3). The relationship was maintained with support moves, albeit in weaker form (r = .24 at Time 2 and $r = .31$ at Time 3, indicating small effect sizes). Regarding speaking anxiety, a generally weak and non-significant negative relationship with topic development moves was maintained throughout ($r = -.14$ at Time 1, $r = -.32$ at Time 2, and $r = -.23$ at Time 3). Concerning support moves, the relationship was neutral at Time 1 ($r = .05$) but became stronger thereafter ($r = -.41$ at Time 2; and $r = -.46$ at Time 3), although it was significant only at Time 3.

Summary of Results

In this chapter I reported the results of the study, which answer the research questions introduced at the end of Chapter 2. The first research question, concerning fluency development, was divided into three parts. First, regarding the monologue activities, significant development was observed in terms of the overall measure of pruned speech rate, as well as the specific measures of mean duration of end-clause silent pauses and mean duration of mid-clause silent pauses. Conversely, the number of repairs per minute significantly increased over time. Next, regarding the group discussions, individuals contributed significantly more fluent syllables over time, whereas the mean length and number of gaps between speaking turns showed large reductions. Finally, the native-speaker ratings showed significant improvements in both fluency and, especially, communicative adequacy during the semester.

The second research question, concerning the development of interactional competence, was analyzed both quantitatively and qualitatively. The quantitative analysis revealed a significant reduction in the number of topic extension statements and minimal extensions produced over time. On the other hand, a significant increase was observed in the number of turn/topic management features produced. No significant change was observed in the number of support moves. The qualitative analysis revealed the emergence of several themes. First, there was a general shift from short—often single or sub-clausal—topic extensions to longer, multi-clausal extensions over time. This trend was matched by a shift from short—often single-word—topic extension questions to longer, more precisely targeted questions. Next, there was a gradual progression towards the smoother use of turn and topic management features, and a reduced reliance on gesture or L1 Japanese during such transitions. Regarding the support features, a gradual diversification of techniques for marking agreement and disagreement was observed over time. Negotiation techniques were also applied to a broader range of situations and there was a reduced reliance on Japanese when negotiating meaning.

Turning to the individual difference variables, the third research question concerned changes in speaking self-efficacy and anxiety over time. Regarding the former, a significant increase was reported between Times 1 and 2 although not between Times 2 and 3. No significant changes were reported regarding speaking anxiety.

The remaining three research questions explored the relationships among the above variables and were therefore answered using correlation analysis. Research question 4 concerned the relationship between fluency and the quantitative aspects of interactional competence. A consistently positive relationship was found between the number of fluent syllables produced during the group discussions and the number of interactional moves—both topic development and support—especially at Times 1 and 2. A positive but weaker relationship between utterance fluency and both types of interactional moves was again more pronounced at Times 1 and 2.

Research question 5 concerned the relationships between the individual difference variables and fluency. A consistently strong and positive relationship existed between extraversion and the number of fluent syllables produced during the group discussions. A positive but weaker relationship was also observed between extraversion and both utterance fluency and perceived fluency. A medium positive relationship was observed between speaking self-efficacy and syllables produced at Time 2 only. In contrast, a negative relationship was observed between speaking anxiety and syllables produced at Times 1 and 2, which disappeared at Time 3.

Research question 6 concerned the relationships between the individual difference variables and interactional competence. A small to medium positive relationship was found between extraversion and the number of topic development and support moves produced at each time point. A consistently positive relationship also existed between speaking self-efficacy and the number of support moves, although with topic development moves this relationship was found only at Time 1. A small to medium negative relationship existed between speaking anxiety and both topic development and support moves at Times 2 and 3 but not at Time 1.

Post-Hoc Analyses

Due to the exploratory nature of this study, the results raised several questions that required further analysis. Therefore, post-hoc analyses were conducted to investigate several of the research questions in more detail. Although these additional results should be interpreted with caution, they also provide some directions for future research.

Research Question 1: Post-Hoc Analysis

Because more pruned speech rate gains occurred between Times 2 and 3, whereas the silent pause gains occurred mainly between Times 1 and 2, I conducted additional analyses to try and locate the source of the gains made in the second half of the course. One possible source is the length of filled pauses, as they were initially analyzed only in terms of rate and location (following De Jong, 2016). Table 62 shows the mean length of all filled pauses (i.e., mid- and end-clause combined) at each time point, presented alongside the mean length of all silent pauses, as well as a composite measure of all pauses for comparison purposes. Associated p-values and effect sizes of the repeated measures ANOVAs are also shown. However, three participants did not use any filled pauses at one of the time points (two at Time 1 and one at Time 3), so they could not be included in this analysis.

Table 62. *Comparison of Total Mean Filled and Silent Pause Length*

	Time 1	Time 2	Time 3	p	η^2
Filled pause mean	2.29	1.77	1.69	.05	.19
Silent pause mean	1.09	0.83	0.82	< .01*	.43
All pauses mean	1.46	1.17	1.13	< .01*	.44

Note. N = 17 for filled pause mean, N = 20 for silent pause mean and all pauses mean.
* Denotes statistically significant result.

It was established previously that mean length of end- and mid-clause silent pauses fell during the course, with gains occurring almost entirely between Times 1 and 2. Unsurprisingly, this finding holds when the two measures are combined into a total silent pause duration measure, as shown in the table above. In this case, a repeated measures ANOVA, $F(2, 38) = 14.58$, $p < .001$, $\eta^2 = .43$, confirmed the significant changes over time, with Bonferroni adjusted post hoc t-tests showing a significant pairwise difference between Times 1 and 3 ($p < .001$, $d = 0.84$) and Times 1 and 2 ($p < .001$, $d = 0.81$, indicating a small effect size in both cases), but not between Times 2 and 3 ($p = 1$, $d = 0.03$).

A repeated measures ANOVA was also conducted to analyze the mean duration of all filled pauses. Mauchly's assumption of sphericity was marginally violated ($p = .047$); therefore, a Greenhouse-Geisser correction was applied, which produced results on the cusp of statistical significance, $F(1.50, 23.97) = 3.67$, $p = .052$, $\eta^2 = .19$. Bonferroni adjusted post hoc *t*-tests were non-significant in each case, despite a small effect size between Time 1 and Time 3 ($p = .06$, $d = 0.58$) indicating a downward trend over time. The pairwise difference between Times 1 and 2 was $p = .11$, $d = 0.51$, and between Times 2 and 3 was $p = 1$, $d = 0.07$.

The combined all pause measure confirms this trend of larger gains in the first half than the second half of the course. In this case, a repeated measures ANOVA, $F(2, 38) = 14.65$, $p < .001$, $\eta^2 = .44$, showed significant changes over time, with Bonferroni adjusted post hoc *t*-tests showing a significant pairwise difference between Times 1 and 3 ($p < .001$, $d = 0.87$, indicating a small to medium effect size) and Times 1 and 2 ($p < .001$, $d = 0.76$, indicating a small effect size), but not between Times 2 and 3 ($p = 1$, $d = 0.12$).

Another potential cause for the pruned speech rate gains that occurred in the second half of the course is a more efficient handling of disfluencies. Table 63 (below) shows the mean number of syllables produced per minute during the monologues at each recording point. There was an average increase of 10.46 total syllables per minute produced between Times 2 and 3, of which 9.73 were fluent syllables (i.e., minus redundancies such as repetitions and false starts). In other words, 93.00% of the additional syllables were fluent. In contrast, of the 8.77 syllables added between Times 1 and 2 on average, only 5.50 were fluent, representing 62.00%.

Table 63. *Mean Number of Syllables Produced per Minute*

	Time 1	Time 2	Time 3
Total syllables	91.70	100.47	110.93
Fluent syllables	79.90	85.37	95.10
Disfluent syllables	11.80	15.10	15.83

Note. N = 20.

Research Question 2: Post-Hoc Analysis

It was established previously that the participants produced fewer topic extension statements over time, whereas the number of fluent syllables produced in the discussions increased. To confirm whether there was a trend towards discussing the topics in greater depth, I additionally analyzed topic extension statements in terms of the total number of words per statement at each time point (see Table 64). In addition, to confirm that any observed patterns were based on meaningful content, the mean number of fluent words (i.e., with redundancies removed) is also presented. Finally, the mean number of fluent words excluding minimal extensions is also shown. Associated p-values and effect sizes of the repeated measures ANOVAs are also indicated.

Table 64. *Mean Number of Words per Topic Extension*

	Time 1	Time 2	Time 3	p	η^2
Total words	15.88	29.19	38.94	< .01*	.34
Fluent words	13.67	25.55	34.26	< .01*	.35
Fluent words minus MEX	16.98	29.93	35.87	< .01*	.32

Note. * Denotes statistically significant result. MEX = minimal extension.

Table 64 shows a clear pattern of increased number of words per topic extension, regardless of whether disfluencies or minimal extensions are included. As the ability to discuss a topic in depth is based on meaningful content, the fluent word statistics are examined in detail below and are used in subsequent analyses. Removing minimal extensions affected the mean much less at Time 3 than at Times 1 and 2, reflecting the fact that sub-clausal extensions were rarely used at this time point, further illustrating the trend towards longer comments. However, they have been retained in the following analysis, as they are a form of topic extension and thus perform a similar, content-generating function.

The descriptive statistics (see Table 65) show the mean increases over time (11.88 words between Times 1 and 2; 8.71 words between Times 2 and 3). However, the large standard deviations and wide confidence intervals indicate a high degree of variance among individuals at each time point. The data were non-normally distributed at each time point, as assessed by the Shapiro-Wilk test ($p < .001$, .01, and .02, respectively), reflecting the presence of outliers (two at Time 1; one at Times 2 and 3). These outliers also contribute to the high kurtosis statistics of between 3.99 and 4.35 at each point. For example, the outlier at Time 3 produced 98.67 fluent words per topic extension, whereas the next highest value was 58.60. However, all three of this participant's comments were relevant to the topic and helped to explore it in depth.

Mauchly's test indicated that the assumption of sphericity was met, with non-significant results for the Chi-square test $\chi^2(2) = 4.82$, $p = .90$. A repeated-measures ANOVA, $F(2, 38) = 10.19$, $p < .001$, $\eta^2 = .35$, confirmed the significant changes over time, with Bonferroni adjusted post hoc t-tests showing a significant pairwise difference between Times 1 and 3 ($p < .001$, $d = 1.30$) and Times 1 and 2 ($p = .04$, $d = 0.75$), but not between Times 2 and 3 ($p = .19$, $d = 0.55$). This finding translates as a relatively large effect size overall, with a small effect in the first half of the course, and a further small effect in the second half.

Table 65. *Descriptive Statistics for Fluent Words per Topic Extension*

	Time 1	Time 2	Time 3
M	13.67	25.55	34.26
SE	2.22	3.37	4.62
95% CI	[9.02, 18.32]	[18.49, 32.61]	[24.60, 43.93]
SD	9.93	15.08	20.65
Skewness	1.96	1.78	1.56
SES	0.51	0.51	0.51
Kurtosis	4.35	4.33	3.99
SEK	0.99	0.99	0.99

Note. $N = 20$ at each time point.

Research Question 4: Post-Hoc Analysis

To further investigate the relationship between fluency and interactional competence, words per topic extension was also correlated with the fluency and communicative adequacy variables. Tables 66 and 67 show no significant relationships between words per topic extension and any other variable at Times 1 and 2, although there were small effect sizes for the correlations with fluent syllables per minute—also based on group discussions—at both time points ($r = .27$ and $.35$, respectively), as well as with the communicative adequacy ratings at Time 1 only ($r = .31$). In contrast, at Time 3 significant relationships were observed between words per topic extension and all other variables (see Table 68). Furthermore, the effect sizes were large for the correlations with syllables per minute and utterance fluency ($r = .67$ and $.58$, respectively) and medium-sized with perceived fluency ($r = .45$). However, the strongest relationship was with communicative adequacy ($r = .72$).

Table 66. *Bivariate Correlations Between Words per Topic Extension Statement and Fluency/Communicative Adequacy (Time 1)*

	1	2	3	4
1. Utterance fluency	—			
2. Fluent syllables per minute	.76***	—		
3. Perceived fluency	.60**	.55**	—	
4. Communicative adequacy	.54**	.72***	.75***	—
5. Words per TES	.28	.27	.02	.31

Note. *** $p < .001$, ** $p < .01$. TES = topic extension.

Table 67. *Bivariate Correlations Between Words per Topic Extension Statement and Fluency/ Communicative Adequacy (Time 2)*

	1	2	3	4
1. Utterance fluency	—			
2. Fluent syllables per minute	.83***	—		
3. Perceived fluency	.71***	.62**	—	
4. Communicative adequacy	.63**	.59**	.70***	—
5. Words per TES	.08	.35	-.01	.03

Note. *** $p < .001$, ** $p < .01$. TES = topic extension.

Table 68. *Bivariate Correlations Between Words per Topic Extension Statement and Fluency/Communicative Adequacy (Time 3)*

	1	2	3	4
1. Utterance fluency	—			
2. Fluent syllables per minute	.68***	—		
3. Perceived fluency	.65**	.59**	—	
4. Communicative adequacy	.48*	.61**	.75***	—
5. Words per TES	.58**	.67***	.45*	.72***

Note. *** $p < .001$, ** $p < .01$, * $p < .05$. TES = topic ext.

Research Question 6: Post-Hoc Analysis

Words per topic extension was additionally correlated with the individual difference variables, to further investigate the relationship between interactional competence and these three variables (see Tables 69 to 71). None of the correlations between words per topic extension and the individual difference variables were statistically significant, due to the small sample size, although some patterns did emerge. First, there was a consistently small relationship between fluent words per topic extension and extraversion (r = .09, .24, and .24, respectively). Its relationship with the other individual difference variables fluctuated markedly, however. The largest changes concerned speaking self-efficacy, which oscillated from r = -.11 at Time 1, up to .43 at Time 2, then back down to -.25 at Time 3. Regarding speaking anxiety, the relationship changed from negative at Times 1 and 2 (r = -.16 and -.34, respectively), to positive at Time 3 (r = .20).

Table 69. *Bivariate Correlations Between Words per Topic Extension Statement and Individual Difference Measures (Time 1)*

	1	2	3	4
1. Words per TES	—			
2. Self-efficacy	-.11	—		
3. Anxiety	-.16	-.16	—	
4. Extraversion	.09	.30	-.31	—

Note. *N* = 20. * *p* < .05. TES = topic extension statement.

Table 70. *Bivariate Correlations Between Words per Topic Extension Statement and Individual Difference Measures (Time 2)*

	1	2	3	4
1. Words per TES	—			
2. Self-efficacy	.43	—		
3. Anxiety	-.34	-.47*	—	
4. Extraversion	.24	.52*	-.32	—

Note. *N* = 20. * *p* < .05. TES = topic extension statement.

Table 71. *Bivariate Correlations Between Words per Topic Extension Statement and Individual Difference Measures (Time 3)*

	1	2	3	4
1. Words per TES	—			
2. Self-efficacy	-.25	—		
3. Anxiety	.20	-.34	—	
4. Extraversion	.24	.25	.08	—

Note. N = 20. * p < .05. TES = topic extension statement.

CHAPTER 6

DISCUSSION

In this chapter, I review and interpret the results of the study. Each research question is addressed in turn, with key findings related to existing research as well as the gaps identified and purposes of the study. Finally, the implications for foreign language learning research and pedagogy are also discussed.

Research Question 1: The Extent of Speaking Fluency Development

Research question 1 concerned the extent to which the student participants' speaking fluency developed during a one-semester course of spoken interaction. Previous longitudinal studies have established that speaking fluency can develop during relatively short time periods in ESL contexts (De Jong & Perfetti, 2011; Mora & Valls-Ferrer, 2012) and study abroad contexts (Freed et al., 2004; Segalowitz & Freed, 2004), especially among higher-proficiency learners. Whether such development can also be observed in lower-proficiency EFL contexts—in which learners typically have limited opportunities to use the target language outside the classroom—has been less clear, although more recent studies have indicated that fluency gains are possible even within a limited time frame (Doe, 2025; Hanzawa, 2021).

A common method of determining whether fluency has developed is to measure aspects of utterance fluency, which are typically subdivided into the three components of speed of delivery, breakdown fluency, and repair fluency (Tavakoli & Skehan, 2005). Changes in measurable outcomes, such as a faster articulation rate or a reduced rate of pausing, imply improved fluency in that specific aspect. More generally, an increase in speech production without an increase in pausing phenomena implies the proceduralization of linguistic knowledge, as speakers are able to produce more speech without the need for more planning time (De Jong & Perfetti, 2011; Towell et al., 1996). Table 4 in Chapter 3 shows the utterance measures used in this study, including an overall measure of pruned speech rate. These measures were based on monologue activities in order to eradicate the potential influence of interlocutors on performance, thus allowing inferences to be made regarding underlying fluency. However, to investigate whether any such gains had translated to the communicative context I also measured aspects of fluency based on performance in group discussions, which has not previously been researched. To this end, the number of fluent syllables produced per minute by each individual was measured, as well as the mean length and number of gaps between turns. Individual performance in the group discussions was also analytically rated to investigate the development of perceived fluency.

Summary of the Results for Fluency Development

Looking at the monologue activities first, significant development was observed regarding the composite measure of pruned speech rate, which encompasses all three aspects of utterance fluency between Time 1 and Time 3. Significant development was also observed regarding the specific measures of mean duration of end-clause silent pauses and mean duration of mid-clause silent pauses, although changes occurred mainly between Times 1 and 2. Conversely, the number of repairs per minute significantly increased over time. All of the other utterance fluency measures remained relatively stable.

Next, regarding the group discussions, individuals produced significantly more fluent syllables over time, with gains occurring between Times 1 and 2 and remaining stable thereafter. The group statistics of number, mean length, and total time of gaps between speaking turns showed large reductions, again specifically between Times 1 and 2. However, these group measures are based on descriptive statistics alone as the groups consisted of different members at each time point.

Finally, the native-speaker ratings showed significant improvements in both perceived fluency and communicative adequacy during the semester. Regarding the former, significance was reached only between Times 1 and 3, and with a relatively modest effect size ($d = 0.59$). Regarding the latter, significant gains also occurred between Times 2 and 3, and with a medium effect size overall ($d = 1.16$).

Interpretation of the Results for Fluency Development

The most relevant comparison with previous research is with those studies that have investigated fluency development in an EFL context. Although it has been claimed that such settings are often unsuitable for the development of oral proficiency, due to the limited quantity and quality of exposure that is often available (Muñoz, 2008), the current study supports the small number of EFL studies that have reported significant fluency gains even within a limited time frame.

Interpretation of Monologue Results

The closest similarities are found with Hanzawa's (2021) study, also based on Japanese freshman university students, and with measures taken from a monologue activity. Although combining silent and filled pauses into one measure, Hanzawa observed that within-clause pause frequency declined after one semester of instruction, from a mean of 1.08 to 0.90 seconds. In the current study, the mid-clause silent pause mean dropped from 0.95 to 0.75 seconds from the beginning to the end of the semester.

Hanzawa also observed a decline in the length of between-clause pauses but only by the end of the academic year, whereas in this study end-clause pauses also dropped significantly within one semester. Similar to this study, Hanzawa found no significant development in terms of articulation rate (expressed inversely as mean syllable duration in this study) or pause frequency.

Overall, there is a clear similarity between the current study and Hanzawa (2021) in terms of the development of breakdown fluency, specifically in the reduction of pause length. It has been hypothesized that clause-end pausing indicates pre-linguistic planning at the conceptualizer stage, whereas mid-clause pausing is associated with linguistic micro-planning, and therefore disfluency at the formulator or articulator stages (Skehan et al., 2016). Hanzawa suggested that these developments reflected the initial stages of fluency development, as a reduction in pause frequency tends to occur at later stages of development and therefore higher proficiency levels (Baker-Smemoe et al., 2014). If so, this interpretation is likely to be even more relevant to the current study given the relatively low proficiency level of the participants. As well as pause frequency, gains in (inverse) articulation rate are also likely to take longer to manifest in an EFL environment with limited opportunities for L2 use outside the classroom. For example, Kahng (2024) found significant syllable duration decreases among intermediate-proficiency participants after five months of study at a university in the United States, as opposed to 14 weeks of EFL instruction as in the current study.

One difference between this study's results and those of Hanzawa (2021) is in terms of the rate of self-repair. Whereas Hanzawa found no difference across the academic year, in the current study it rose from 2.20 instances per minute at Time 1 to

3.63 at Time 3 ($p = .02$, $d = 0.88$). This finding might seem counterintuitive given that an increase in the rate of repair appears to signify a decrease in fluency. However, it is congruent with studies that have found repair rate increases among lower-proficiency learners and, conversely, decreases among higher-proficiency learners (e.g., Leonard & Shea, 2017). As the ability to self-repair depends on the effective monitoring of one's own output, a possible explanation for such findings is that increased oral proficiency and metalinguistic awareness allows more attention to be diverted to the monitoring process, thus leading to an increase in self-repair frequency (Kormos, 1999). In other words, an increase in the rate of repair at lower proficiency levels—such as in the current study—might imply more efficient monitoring processes resulting from increased linguistic knowledge. This account could also explain rate of repair decreases at more advanced levels, as once a threshold of accuracy and efficiency in linguistic encoding has been reached the need for overt self-repair is reduced (Lambert et al., 2017), resulting in what Hanzawa (2021) has called "an inverse U-shaped trajectory" (p. 17). Furthermore, Hanzawa found a correlation between increased self-repair and the amount of study reported by the participants in both semesters. This finding suggests that students who spent more time learning English were more likely to increase the number of self-repairs over time. Similarly, in this study the fact that repair frequency increased steadily over the semester implies a gradual increase in the ability to self-monitor as linguistic knowledge was accumulated.

There are also many contextual parallels between the current study and Doe's (2025) study, which similarly used monologue activities to investigate fluency gains within one semester of Japanese university discussion-based speaking classes. Doe reported small but significant gains in phonation/time ratio—used to calculate the amount of pausing in a speech sample—and mean length of pause (from 1.35 to 1.16 seconds), although he did not distinguish between type or location of pause. The fact that Doe did not observe gains in other fluency measures, including articulation rate, implies that his study also captured the initial stages of fluency development.

Several other longitudinal studies of fluency development have used EFL classes as control groups to compare with ESL or study abroad contexts (Freed et al., 2004; Mora & Valls- Ferrer, 2012; Segalowitz & Freed, 2004). The participants in the current study made substantially more fluency gains than those EFL control group students, although in most cases the latter appear to have had limited speaking opportunities. It is therefore unsurprising that fewer gains were made when compared with students in highly communicative environments, whether at home—as in the current study—or abroad. Taken as a whole, these findings support the need for extensive practice to proceduralize, and eventually automatize, linguistic knowledge, as stated in DeKeyser's (2007) application of skill acquisition theory to SLA.

In sum, the monologue fluency measures showed gains in utterance fluency, as significant development was observed regarding mean length of silent pause, both mid-clause and at clause-end boundaries. Furthermore, the fact that the silent pause gains occurred almost entirely in the first half of the course implies that they were probably not related to underlying cognitive fluency, that is, the efficiency with which speakers integrate the mental processes underpinning speech production, at this stage. This interpretation is supported by De Jong et al.'s (2013) assertion that silent pause duration is a weak indicator of cognitive fluency, based on their finding that it had the lowest relationship to linguistic knowledge and processing skills, explaining only 5.00% of the variance.

The mean length of silent pause gains are also reflected in the overall measure of pruned speech rate, which showed a significant increase over the semester despite an increase in the number of repairs. However, more speech rate gains occurred between Times 2 and 3, whereas the silent pause gains occurred mainly between Times 1 and 2. The post-hoc analysis of filled pause duration (see Table 62) also revealed a larger decrease between Times 1 and 2 than between Times 2 and 3, indicating that this variable could not be responsible for the gains in pruned speech rate that occurred in the second half of the course.

In addition, it is noticeable from Table 62 that the mean duration of filled pauses was more than double that of silent pauses at each time point. It has been argued that filled and silent pauses are distinct as the former have communicative value in terms of maintaining the interactional flow (Peltonen, 2017) and signaling the speaker's desire to continue (De Jong, 2018). However, the filled pauses in this study were often extended by silence on either side of the vocalization, which is likely to be the case at lower proficiency levels. In such cases the distinction between silent and filled pauses becomes blurred, as the latter do not only consist of the vocalized element.

Furthermore, it is possible that filled and silent pauses are caused by similar cognitive issues in the speech planning process, but that some speakers use filled pauses whereas others use silent pauses to stall for time when encountering such problems (De Jong et al., 2013). Table 72 (below) shows each individual's rate of pausing, combined across the three time points, ranked by ratio of total silent to filled pauses shown in the end column. Although unfilled pauses have been found to negatively correlate with speaking proficiency (Iwashita et al., 2008), these results appear to support De Jong et al.'s suggestion that individuals often use different pausing behaviors to deal with similar cognitive processing issues. For example, Naoya had the highest rate of total silent pauses per minute (26.89) but the lowest rate of total filled pauses (0.44), and Airi had the third-highest rate of total silent pauses (18.66) but the second lowest rate of total filled pauses (1.56). Conversely, Kenji had the second highest rate of total filled pauses (16.22), but the lowest rate of total silent pauses (7.55). Such dramatic differences in pausing behavior are unlikely to reflect differences in speaking proficiency given that all participants in this study had been placed in the same proficiency level.

Moreover, pausing behaviors do not always remain stable. For example, Tetsuya switched from a reliance on silent pauses in the first half of the course to filled pauses at the end. This change was especially pronounced in terms of mid-clause pauses, when learners are most likely to be formulating their linguistic message. At Time 1 his rate of silent and filled mid-clause pauses was 15.33 and 5.33, respectively, which at Time 2 became 21.33 and 4.67. However, at Time 3 this tendency had reversed such that his rate of silent mid-clause pauses was 4 whereas for filled mid-clause pauses it was 10. Therefore, the fact that learners adopt differential pausing behaviors, and not always consistently, further blurs the distinction between silent and filled pauses. Although these results are based on monologue activities, perhaps the main distinction lies in how such pauses are perceived by listeners. As De Jong (2018) pointed out, a filled pause has greater communicative utility because it provides a more overt signal that the speaker intends to hold the floor, thus providing information that benefits the listener.

Table 72. *Individual Rate of Total Filled and Silent Pauses*

Part.	SP mid	SP end	SP total	FP mid	FP end	FP total	SP/FP
Kenji	6.22	1.33	7.55	12.00	4.22	16.22	0.47
Manabu	10.89	1.78	12.67	12.44	6.67	19.11	0.66
Midori	9.55	2.67	12.22	8.89	4.67	13.56	0.90
Kiyomi	6.45	4.44	10.89	5.78	3.33	9.11	1.20
Katsuya	11.33	5.56	16.89	9.11	4.22	13.33	1.27
Keita	13.78	3.33	17.11	8.00	3.56	11.56	1.48
Yukino	9.33	5.78	15.11	8.00	1.11	9.11	1.66
Kaito	10.22	2.22	12.44	4.89	1.78	6.67	1.87
Rie	12.00	3.78	15.78	6.44	2.00	8.44	1.87
Tetsuya	13.55	4.00	17.55	6.67	1.78	8.45	2.08
Naho	15.33	2.89	18.22	4.22	2.45	6.67	2.73
Ryota	12.22	3.11	15.33	4.67	0.89	5.56	2.76
Akari	15.33	6.89	22.22	7.11	0.22	7.33	3.03
Konomi	15.11	3.56	18.67	3.11	2.00	5.11	3.65
Aoi	17.34	6.67	24.01	4.22	0.67	4.89	4.91
Emiri	13.78	6.67	20.45	1.78	2.22	4.00	5.11
Honoka	13.33	5.33	18.66	2.00	0.45	2.45	7.63
Shinpei	11.78	6.45	18.23	1.11	0.89	2.00	9.12
Airi	16.66	5.11	21.77	0.89	0.67	1.56	13.96
Naoya	18.67	8.22	26.89	0.22	0.22	0.44	61.11

Note. Part. = participant, SP = silent pause, FP = filled pause, mid = mid-clause, end = end-clause.

Another potential cause for the pruned speech rate gains that occurred in the second half of the course is a more efficient handling of disfluencies. Table 63 showed that, although the number of repetitions and repairs did not change significantly between Times 2 and 3, the proportion of fluent syllables did increase. This change therefore contributed to the higher pruned speech rate.

One potentially fluency enhancing feature that these gains do not appear related to is the use of the formulaic phrases introduced during the course, such as *In my opinion*. Although formulaic language has been found to correlate positively with speech and articulation rate (Tavakoli & Uchihara, 2019; Yan, 2020), mean syllable duration in the current study was the same at Time 3 as at Time 1, so their use did not result in a faster speaking speed. Moreover, many of these phrases relate to turn-taking and managing interaction (e.g., *Can I start?*) and were therefore not relevant to the monologue activities on which the utterance fluency measures were based. Although these formulaic phrases, even if automatized, did not constitute a high enough proportion of total output to impact the monologue fluency results, it is possible that they influenced fluency during the group discussions—reflected in the increased number of fluent syllables produced per minute—as the phrases related to turn-taking and interaction could also be used in this context.

To conclude, the gains in pruned speech rate in the first half of the course were driven mainly by reduced pausing time, which is a sign of incipient fluency development (Hanzawa, 2021). However, development in the second half of the course was due to an accumulation of marginal gains across various utterance fluency measures, none of which reached statistical significance. For example, the rate of mid- and end-clause pauses fell between Times 2 and 3, having both increased between Times 1 and 2 (see Table 33), reflecting the kind of non-linear patterns associated with L2 fluency development (De Bot & Larsen-Freeman, 2011; Tavakoli et al., 2020).

Development in the latter part of the course was also due in part to learners becoming more efficient at dealing with disfluencies as they increased their linguistic knowledge and monitoring capabilities. Furthermore, the fact that more fluent syllables were produced while pausing phenomena remained stable at this point implies that proceduralization had occurred (De Jong & Perfetti, 2011; Towell et al., 1996). In other words, cognitive restructuring might be underway, as the various aspects of speech production began to work more efficiently and in concert (Segalowitz, 2010).

Interpretation of Group Discussion Results

Having established that fluency gains were manifested in monologue tasks, the next step is to investigate whether such gains translated to the interactive context. However, there are no studies to compare these results with, as to the best of my knowledge fluency development in a group discussion context has not previously been researched. Table 63 (above) showed that the participants produced more fluent syllables in the monologues over time. Similarly, Table 37 in Chapter 5 showed that the participants produced more fluent syllables in the group discussions throughout the course (30.75 per minute at Time 1, 38.48 at Time 2, and 39.73 at Time 3). These findings imply that learners were able to transfer this aspect of fluency, namely the ability to produce more speech within a fixed time frame, from the individual to the communicative context. The main difference between the two findings is that, whereas the total syllable gains were spread relatively evenly in the monologue context (8.77 per minute between Times 1 and 2; 10.46 between Times 2 and 3), in the group context the gains were observed mainly in the first half of the course (7.73 per minute between Times 1 and 2; 1.25 between Times 2 and 3). This discrepancy suggests that a ceiling was reached relatively quickly during the group discussions, perhaps due to the limitations on speaking time imposed by the group format and the requirements of turn taking.

The increased number of syllables relates directly to the other group discussion fluency measures, as the scope for increased talking time was produced, at least in part, by a reduction in the number and mean length of gaps between turns. Specifically, there was a large reduction (11.92) in the number of gaps of at least one second between Times 1 and 2, and a slight increase (1.80) between Times 2 and 3. In addition, there was a larger reduction in the mean length of such gaps (0.62 seconds) between Times 1 and 2, and a smaller reduction (0.31 seconds) between Times 2 and 3. These two findings were corroborated by a very large reduction in the total time of inter-turn gaps of 55.20 seconds between Times 1 and 2, but only 3.10 seconds between Times 2 and 3. It is possible that a reduction in intra-turn pauses was further responsible for the increased number of syllables produced; however, as mentioned in Chapter 3, such pauses were not measured in this study due to the difficulties of doing so reliably in the presence of interlocutors.

Taken together, the discussion-based measures provide further evidence of fluency development made during this one-semester course. In the monologue activities, gains took the form of an increase in pruned speech rate—driven early on by a reduction in pausing time—which was reflected in the higher number of syllables produced. Similarly, in the discussion activities participants were able to increase the number of syllables they produced, especially in the early part of the course. In this case, the gains were at least partly driven by a reduction in inter-turn pausing time. This finding also has implications for listeners' perceptions of fluency, as Van Os et al. (2020) found that extended gaps between turns caused nonnative speakers to be rated negatively. Furthermore, in terms of the mean length of gaps between turns, confidence intervals narrowed and standard deviations fell throughout the course. The fact that variance between groups fell after the first recording implies that the changing discussion groups, and ultimately the members they consisted of, gradually converged on a shared style of interaction that enabled more content to be produced.

It has been argued that fluency in interaction represents a different construct from that displayed during a monologue, due to the demands of accommodating interlocutors in the former context (Tavakoli, 2016). Furthermore, fluency in interaction is largely co-constructed, given the intertwined nature of speaker and listener roles (Kasper, 2006), and the requirement of speakers to relate their comments to previous turns (McCarthy, 2010). Nevertheless, the similarity of findings across both contexts in this study, in terms of speech production and pausing phenomena, imply an overlap between the two. These findings also suggest that learners can transfer gains made in one context to the other.

Interpretation of Ratings Results

Having observed fluency gains in both the monologue and group discussion contexts, the final step is to investigate whether this development is noticeable by listeners. Previous research has established a consistent connection between pausing phenomena and speed of delivery on the one hand, and raters' perceptions of fluency on the other (De Jong et al., 2013; Suzuki et al., 2021). In other words, the faster learners speak and the less they pause the likelier they are to be rated as fluent—a finding that holds regardless of the raters' background (i.e., expert or novice, trained or untrained, native or non-native speaker). For this reason, the analytic rating scale used in the current study referred to speed and hesitation in each level descriptor (see Appendix R).

In this study, a small but significant improvement was observed by raters across the semester. Although statistical significance was reached only between Times 1 and 3 (p = .04, d = 0.59), average ratings increased more between Times 2 and 3 (p = .17, d = 0.49) than between Times 1 and 2 (p = 1, d = 0.14). This latter result contrasts with the gains made in terms of fluent syllables produced in the group discussions, which occurred mainly in the first half of the course. Furthermore, it is unlikely to reflect the gains regarding gaps between turns, which also occurred mainly between Times 1 and 2. Moreover, the ratings were applied to individual participants, and such gaps by definition cannot be assigned to individuals. It is therefore hard to know what specifically prompted the raters to assign higher fluency scores in the latter part of the course, especially as there is no guarantee that what appears salient to raters actually reflects an individual's underlying cognitive fluency (De Jong, 2018).

A clue as to why the perceived fluency ratings might have improved more between Times 2 and 3 is provided by the communicative adequacy ratings, which were included to confirm that the goals of the task had been achieved. For this scale, the descriptors referred to the comprehensibility and amount of relevant detail included in the participants' contributions (see Appendix R). These ratings improved significantly not only between Times 1 and 3 ($p < .001$, $d = 1.16$), but also between Times 2 and 3 ($p < .001$, $d = 0.87$), although not between Times 1 and 2 ($p = .46$, $d = 0.28$), indicating that substantial development took place in the second half of the course. These results, therefore, followed a similar—but more pronounced—pattern than those for perceived fluency, with larger effect sizes at each time point. Révész et al. (2016), using broadly similar descriptors to those used in the current study, found that utterance fluency predicted communicative adequacy, in a variety of communicative situations, more strongly than accuracy or complexity. However, the reverse could also be true in that communicative adequacy could impact on how fluent a sample of speech is judged to be. Therefore, the fact that the raters scored the participants considerably higher for adequacy (i.e., how detailed and comprehensible their contributions were) in the second half of the course could have influenced how fluent they felt those contributions to be. If so, this finding suggests some overlap between how the two constructs were perceived. It could also be influenced by the fact that raters are often reluctant to assign markedly different scores to the same individual (Ockey, 2011), which is also known as the halo effect in Rasch analysis.

It was established in Chapter 5 that the three individual fluency measures—utterance fluency (i.e., pruned speech rate), fluent syllables per minute (based on group discussions), and perceived fluency—were all positively and significantly correlated at each time point, with large effect sizes produced in almost every case. In other words, participants who performed more fluently in the monologues produced more speech during the discussions and were rated as more fluent in those discussions. This finding implies that all three measures were tapping into different aspects of the same basic construct of L2 speaking fluency. Furthermore, the strongest of these correlations was consistently between utterance fluency (i.e., based on monologues) and syllables per minute (i.e., based on group discussions), indicating that learners with the highest underlying fluency—as inferred from monologue delivery—tended to talk the most in the interactive context. The communicative adequacy ratings maintained this pattern of consistently strong, positive relationships with each variable. Moreover, the fact that perceived fluency had the strongest correlation with communicative adequacy at each point further implies some overlap between how these two constructs were perceived by the raters, and that they were perhaps influenced by one when rating the other.

The fact that significant gains were made in all the aspects of speaking fluency addressed in this study suggests that L2 learners can make substantial progress in an EFL environment, even with just a single semester of study. However, the context of the study is of vital importance. This course consisted of communicative speaking classes, one of which was discussion-based, with students expected to use English only in class. Many opportunities were provided for meaning-focused output, and repeated monologue activities were also regularly used to further develop fluency. Furthermore, classes were held twice a week, in contrast to the standard once a week classes that are held in most departments in most Japanese universities. In addition, the students took two other twice-weekly, required English courses—a listening and a reading and writing course—which provided further opportunities for engaging with the target language. This learning context, therefore, provided the opportunities and conditions necessary for fluency development to take place.

Research Question 2: The Extent of Interactional Competence Development

Research question 2 concerned the extent to which the participants' interactional competence developed during a one-semester course of spoken interaction. Most interactional competence research to date has been conducted qualitatively, with classroom-based researchers focusing on the development of individual micro-skills such as repair or disagreement techniques. Such research has established that interactional competence development includes: taking the initiative during interaction (Young & Miller, 2004); engaging in collaborative turn sequences (Sert, 2019); and expanding the range of techniques that learners are able to utilize (Pekarek Doehler & Pochen-Berger, 2015).

Nevertheless, some attempts have been made to quantify the features of L2 interaction, especially in terms of distinguishing between proficiency levels. For example, Galaczi (2014) found that higher-level learners displayed more mutuality (i.e., engagement with each other's contributions) and were more collaborative than their lower proficiency peers, as they developed other-initiated topics, offered more listener support, and demonstrated more efficient turn-taking. However, such research has focused mainly on high-stakes assessment, with coding schemes tailored to the tasks and kinds of language produced in those contexts. Whether quantification can also illustrate longitudinal development in the types of unassessed group discussions that commonly occur in EFL classrooms has not previously been investigated. Therefore, in order to quantify the interactional features used in the current study, Galaczi's (2014) coding system was modified to fit the current context. As well as for quantification purposes, this coding system was then used as the basis for the subsequent qualitative analysis.

Summary of the Results for Interactional Competence Development

Beginning with the quantitative results, non-parametric analysis showed a significant reduction in the number of topic extension statements and minimal extensions produced over time. In the former case, the reduction was statistically significant only between Time 1 and Time 3, whereas in the latter case significance was reached between Times 1 and 2 as well as Times 1 and 3. In addition, there was a relatively large drop in the number of topic extension questions in the early part of the course, albeit due to the behavior of a small number of participants and therefore not statistically significant. In contrast, the number of turn/topic management moves significantly increased over time. Significance in this case was reached between Times 1 and 2 as well as Times 1 and 3. No significant change was observed in the number of support moves, which all remained stable over time. Post-hoc analysis revealed a significant increase in the number of words per topic extension statement between Times 1 and 2, as well as between Times 1 and 3.

Regarding the qualitative analysis, several themes emerged from the data. In terms of topic development moves, there was a shift from short or minimal topic extension statements to longer, multi-clausal extensions, which was matched by a shift from short or lexical topic extension questions to longer, more precisely targeted questions. There was also a gradual progression towards the smoother use of turn and topic management features, as well as a reduced reliance on gesture or Japanese use. Regarding the support features, there was a gradual diversification of methods for marking (dis)agreement and meaning negotiation, which were utilized across a broader range of situations.

Interpretation of the Quantitative Results for Interactional Competence Development

As mentioned above, the coding system used in the current study was based on Galaczi's (2014) system. However, it was modified due to some important contextual differences, as high-stakes paired speaking exams and unassessed group EFL discussions are fundamentally different types of interaction (see Chapter 3 for a fuller explanation of the differences between the two contexts and a justification for the system used in this study—Tables 6 and 7 provide a full list of the topic development and support moves included). Nevertheless, as with almost any form of interaction, both contexts share a common need to generate content and maintain comprehension; therefore, it is instructive to compare findings to ascertain whether development can be observed in the current study.

One aspect of collaborative interaction that Galaczi (2014) identified among higher-proficiency learners was turn-taking efficiency. She argued that gaps between turns result from the increased processing time that lower-proficiency speakers need to decode the previous turn and then construct a response. Other research has also identified inter-turn gaps as a feature of lower-level interaction (e.g., Galaczi, 2008; Sert, 2019). As mentioned previously, a reduced number and length of gaps between turns was observed in the current study, although this finding was included in the fluency section due to its temporal nature.

Regarding topic development, Galaczi (2014) found that higher-proficiency test takers developed topics initiated by other speakers much more than their lower-proficiency peers. She attributed this finding to the increased cognitive burden required to simultaneously maintain speaker and listener roles, which is necessary when developing other-initiated topics. In the current study, however, the burden of topic initiation was largely removed from the participants by the provision of a limited number of discussion prompts, implying the need to discuss a topic in depth rather than continually initiate new ones. Therefore, the fact that the participants produced fewer topic extensions over time, whereas the number of fluent syllables produced in the discussions increased (as discussed in the previous section), suggests that they were contributing fewer but longer comments.

This suggestion of longer comments was confirmed by the post-hoc analysis of words per topic extension. However, there was a wide degree of variance between individuals, reflecting the contingencies of interaction, as shorter comments and responses are sometimes appropriate even in a discussion, such as in response to a closed follow-up question. Nevertheless, it also reflects a wide disparity among individuals in terms of their propensity to provide fuller contributions.

Table 73 (below) shows the number of fluent words per extension produced by each individual. The data are ranked by length of topic extension statement at Time 3, with the number of topic extension statements (including minimal extensions) in parentheses. This table confirms the variation between individuals. For example, Keita's contributions were consistently among the longest, whereas Kiyomi's were consistently among the shortest. Even so, Kiyomi roughly tripled the length of her extensions at Time 2, which had then quadrupled by Time 3, illustrating that the participants' development occurred relevant to their starting base. Furthermore, Kaito, Manabu, Naho, and Tetsuya all saw large increases over time from a low starting base.

Overall, despite the large amount of variance at each time point, there was a clear trend towards longer topic extensions throughout the semester, and therefore more detailed comments during the discussions. Such comments indicate more complex output, as length of speaking turn—in this case topic extension statement—is also a measure of interactional complexity (Ellis & Barkhuizen, 2005). Furthermore, supporting opinions with reasons and examples—as called for in discussion activities—also implies greater propositional complexity, in which more idea units are included in the speaker's turn (Ellis & Barkhuizen, 2005). Producing longer extensions also has implications for fluency development, as such comments provide more of the practice that speakers need to proceduralize linguistic knowledge (DeKeyser, 2007).

Table 73. *Individual Rate of Fluent Words per Topic Extension*

Participant	Time 1	Time 2	Time 3
Aoi	21.08 (12)	26.38 (8)	98.67 (3)
Emiri	12.00 (15)	12.53 (20)	58.60 (5)
Manabu	6.53 (15)	19.73 (11)	52.00 (4)
Keita	33.40 (5)	45.33 (6)	46.40 (5)
Katsuya	16.93 (15)	38.44 (9)	43.67 (6)
Yukino	44.25 (4)	16.78 (10)	41.17 (5)
Midori	2.27 (15)	20.40 (19)	39.67 (6)
Naho	9.78 (17)	35.86 (6)	38.57 (7)
Konomi	13.80 (10)	18.50 (10)	36.33 (3)
Ryota	14.50 (5)	27.00 (4)	33.00 (4)
Kaito	6.25 (4)	13.09 (11)	30.50 (6)
Tetsuya	6.15 (20)	30.75 (4)	30.40 (5)
Airi	9.22 (10)	34.20 (5)	27.60 (5)
Kenji	14.11 (9)	23.00 (8)	23.80 (5)
Akari	14.30 (10)	73.33 (3)	19.55 (10)
Rie	5.19 (16)	27.17 (6)	17.00 (6)
Shinpei	15.57 (7)	9.64 (11)	16.57 (7)
Naoya	4.55 (23)	15.50 (6)	12.21 (14)
Kiyomi	3.06 (17)	9.36 (11)	12.20 (5)
Honoka	10.50 (10)	14.00 (6)	7.38 (8)

Moreover, this trend indicates that learners had adapted to the local demands of the task, thus demonstrating an ability to deal with a wider range of interactional contexts, which has been described as the key to interactional competence development (Pekarek Doehler & Pochen-Berger, 2015). For example, an extended discussion in an academic environment implies a need for longer, more complex comments than a casual greeting in a social environment. Arguably, therefore, words per topic extension is a better measure of development in this context than number of topic extensions, as it implies engaging with the topic in depth. Consequently, the trend towards longer topic extensions in this specific context represents the development of this aspect of interactional competence.

This trend towards longer topic extensions also brought the student participants closer to the performance of the three highly proficient Japanese speakers of English.

These speakers, who had one group discussion separately recorded, produced an average of 37.67 fluent words per topic extension—only 3.41 above the student participants at Time 3. To illustrate this point, one of Aoi's (who produced the longest topic extensions at Time 3) comments is reproduced below, together with a typical comment from the highly proficient speaker discussion (redundancies removed in both cases) for comparison purposes. First, Aoi's comment:

> I think in present situation high school life is better. That is because that we don't have any time to consider about the university life. However, I think that university life is so freedom to decide what we do. However, we have a lot of responsible. For example, that we are absent from the class it is okay but finally we have a lot of burden, so I think that it is sometimes better, sometimes it's not so good.

Next, the highly proficient speaker comment:

> I also think nature is really beautiful in Japan, so my favorite place is Izu peninsula in Shizuoka because it's a peninsula so you can go to the beach, and actually there's a cliff and really rugged coastline. But also you can go to the mountainside, and you can see beautiful scenery from the mountain top so it's a lot of things to see in Izu.

Unsurprisingly, the highly proficient speaker's comment is more grammatically accurate and includes more low-frequency vocabulary (e.g., *rugged* and *peninsula*) than the low-intermediate student comment shown above. However, it reveals a similar inclination to discuss the topic in depth, with several reasons and examples provided to support the opinion. The highly proficient speakers were given no instruction regarding how to conduct their discussion; therefore, if the goal of learners is to approximate the natural behavior of such speakers, the fact that they progressed towards this behavior throughout the course implies development in this aspect of their performance. A reduction in the number of minimal extensions over time, to 0.90 at Time 3, also brought the student participants closer to the behavior of the highly proficient speakers, who produced an average of 0.67 minimal extensions, emphasizing their inclination to produce extended comments when presented with a discussion task.

Despite this convergence, it should be emphasized that longer comments do not necessarily equate to better interaction, even in this context. For example, according to Sert (2019), extended individual turns characterized novice performance in EFL group discussions conducted among Turkish university students. However, he also reported concurrently long inter-turn gaps and a lack of listener support (e.g., acknowledgement and agree/disagree tokens) as indicative of such performance. In the current study, the participants were able to offer such support during—and in response to—extended comments, as shown by the excerpts in Chapter 5. Moreover, they did so to a degree that was quantifiably comparable with the highly proficient speakers, as discussed below. Also, as mentioned previously, inter-turn gaps decreased at the same time as topic extensions expanded. Therefore, I argue that extended comments in an academic group discussion context can indicate development—and coexist with other forms of development—especially when compared with the kind of superficial interaction shown in Excerpts 2 and 5 in Chapter 5. Furthermore, the fact that the raters evaluated communicative adequacy highest at Time 3—when the topic extensions were at their longest—supports this claim, given that the 5-point descriptor refers to "contributions… with appropriate amount of detail."

Another potential counter to this argument is that longer extensions could reflect dominance by individual participants at the expense of other members, which Galaczi (2008) referred to as asymmetric patterns of interaction. Although this outcome is certainly possible, the data in this study do not indicate that such dominance occurred.

Table 73 shows the number of topic extension statements that each individual made next to the average number of words per extension. At Time 3, when the extensions were longest on average, none of the four highest ranked participants made more than five statements. In addition, only three participants in total made more than eight or more statements, all of whom were ranked in the bottom six in terms of words per statement. Furthermore, correlation analysis revealed large and significant negative relationships between number of extensions and number of words per extension at each time point (r = -.60, p = .005 at Time 1; r = -.54, p = .02 at Time 2; r = -.55, p = .01 at Time 3). The scatter plots shown in Figures 14 to 16 (below) visually represent these relationships. The lack of any values in the top right quadrants illustrates the fact that no participant produced a large number of long extensions at any time point.

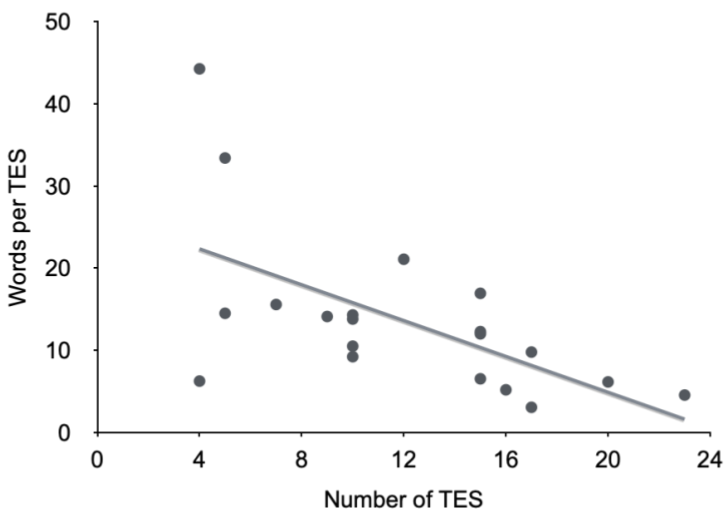

Figure 14. Scatterplot of Number and Words per Topic Extension Statement (Time 1)
Note. TES = topic extension statement.

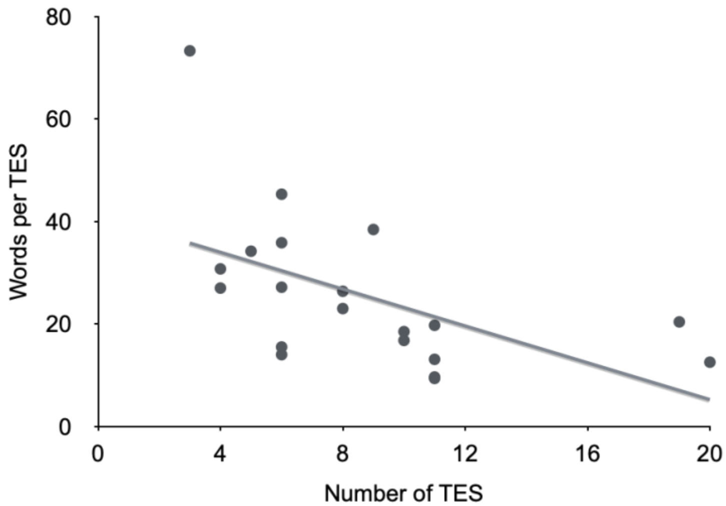

Figure 15. Scatterplot of Number and Words per Topic Extension Statement (Time 2)
Note. TES = topic extension statement.

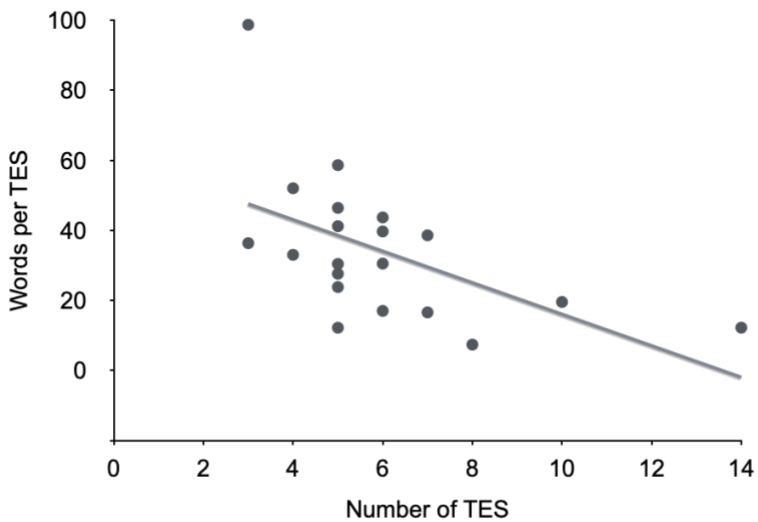

Figure 16. Scatterplot of Number and Words per Topic Extension Statement (Time 3)
Note. TES = topic extension statement.

Regarding support moves, Galaczi (2014) found that higher-proficiency test takers offered more listener support, in the form of backchannels and confirmations of comprehension, than their lower-proficiency peers. Galaczi again attributed this result to the difficulty lower-level learners have in simultaneously maintaining speaker and listener roles. In the current study, no quantifiable change was observed regarding any of the support features, which includes repair and agree/disagree markers in addition to confirmations and backchannels. It should be noted, however, that the number of such features produced was broadly comparable to that produced by the highly proficient speakers, as shown in Table 51 in Chapter 5. Using the end of course (i.e., Time 3) measures for the student participants, the proficient speakers produced more backchannels (16.00 vs. 11.05) and agree/disagree markers (1.33 vs. 0.95), implying slightly more engagement with the speaker. On the other hand, they produced fewer confirmations (1.00 vs. 2.10) and repairs (0.00 vs. 0.95), which can be explained by less need to negotiate meaning among highly-proficient speakers. Nevertheless, the fact that the student participants were able to consistently produce these support features throughout the course implies that limited cognitive processing capacity was not a barrier to this kind of collaborative interaction, which supports Waring's (2019) assertion that low linguistic proficiency does not automatically imply low interactional competence.

The only interactional feature that increased quantifiably in the current study was the category of turn/topic management, with such moves increasing significantly between Times 1 and 2, as well as between Times 1 and 3. This finding tallies with the decrease in number and length of gaps between turns, which also occurred mainly in the first half of the course. The fact that these increases in collaborative behavior occurred early in the course also suggests that they might not have resulted from increased cognitive processing capacity, which would likely take longer to manifest, as discussed in the fluency section above. Rather than cognitive development, reduced gaps between turns could reflect changes in affect, such as increased self-efficacy, which are discussed later in this chapter. In addition, the increased use of turn and topic management features is probably related to their pedagogical introduction in the first half of the course, which is also discussed below.

To conclude, quantifying the topic development and support moves produced during the group discussions highlighted the development of some aspects of interactional competence. First, the participants produced fewer but longer topic extensions as they adapted to the demands of the task. Although Galaczi's (2014) distinction between *own* and *other* topic extension was rendered moot by the provision of discussion questions in advance, the fact that the participants moved towards longer topic extensions over time shows progression in terms of "degree of topic development," which Galazci identified as one of the most salient differences between higher and lower speaking proficiency levels.

There was also a significant increase in the number of turn/topic management moves which, when combined with the reduction in inter-turn gaps, indicates a trend towards more collaborative behavior. Furthermore, the occurrence of these gains mainly in the first half of the course, combined with the fact that support moves were produced consistently throughout the course, suggests that even relatively low-level students can overcome cognitive processing limitations from the earliest stages when contributing to multi-party interaction. In other words, learners can draw on the universal communication abilities that precede L2 learning—or even L1 learning—to establish collaborative patterns of interaction (Hall, 2018; Levinson, 2006). As fluency develops, learners can theoretically build on this base and use the increased processing capacity to contribute longer and more complex comments to the discussion.

Interpretation of the Qualitative Results for Interactional Competence Development

The quantitative findings are supplemented by three main themes that emerged from the qualitative data. The first theme relates to the length and detail of topic development moves. Quantification showed that learners refined their use of topic extension moves throughout the course, progressing towards longer, multi-clausal comments rather than a series of short or sub-clausal minimal extensions. Qualitative examples of this process were shown in Excerpts 2 to 6 in Chapter 5. This shift had repercussions in terms of fluency and affective development, which are explained in the relevant sections of this chapter. Similarly, topic extension questions became more detailed and precisely targeted over time, as shown in Excerpts 7 to 10.

In terms of interactional competence development, adapting to the tasks in this way signaled an increased ability to deal with a range of interactional contexts, given that the background questionnaire revealed group discussion activities were unfamiliar to these students prior to entering university. This finding also highlights the contingent nature of interactional competence, as not all forms of interaction call for extended turns. For example, Galaczi (2008) stated that short turns were indicative of a high-involvement, collaborative style of interaction based on mutual topic development. However, her study was based on a high-stakes paired speaking exam in which picture prompts provided many options and opportunities to initiate new topics. In contrast, a group discussion activity with only two question prompts based on one pre-determined topic calls for participants to discuss a topic in depth by giving and supporting opinions, which were the foci of Weeks 2 and 3 of the course (see Table 1 in Chapter 3 for the course schedule and Appendix E for a list of all the discussion skills and formulaic phrases that were taught). Furthermore, the unassessed nature of the activity put less pressure on the participants to compete for speaking turns in order to display their competence. Nevertheless, listener support can still be provided while a topic extension is being produced, as was often the case in the current study—exemplified by the backchannels in Excerpts 4 and 6.

The second theme to emerge from the qualitative data was a gradual progression towards the smoother use of turn and topic management features. Quantitative analysis revealed that this was the only feature to increase significantly over time, indicating more collaborative behavior as the participants took responsibility for transitioning between turns and topics. It also implies a greater willingness to take the initiative, which has been identified as a key aspect of interactional competence as lower-level learners tend to be more passive and often wait to be invited to participate (Pallotti, 2019).

The qualitative data showed that at Time 1 there were often extended inter-turn pauses (see Excerpts 2, 7, and 29). In addition, the participants often relied on L1 Japanese or non-verbal means when transitioning between turns and topics, leading to some awkward exchanges (see Excerpt 11). This tendency was reduced at Time 2, however, as several participants attempted to use the *Joining a Discussion* formulaic language (e.g., *Can I start?*) that that had been introduced in Week 5 of the course, prior to the second recording. Although using such language involves taking the initiative in terms of managing the discussion, Excerpts 12 to 14 show that it was not always used accurately or fluently at this stage, suggesting that learners needed time to practice such phrases in order to use them effectively. At Time 3, however, these phrases and questions were generally used more accurately and were often combined in collaborative transition sequences (see Excerpt 15). Turn-taking dynamics are a fundamental element of interactional competence (Pallotti, 2019); therefore, an improved ability to transition smoothly between turns and topics, without relying on the use of gestures or the L1, is another indicator of development. Examining the use of turn/topic management features in this study also illustrates the need for both qualitative and quantitative analysis, as they revealed different aspects of development. Furthermore, smoother transition sequences are likely related to the reduced gaps between turns discussed in the previous section.

The third qualitative theme relates to the use of support features. Quantification showed that support moves, in all categories, were contributed from the earliest stages of the course and that their use remained relatively stable in numerical terms. In this case, development was observed solely through qualitative analysis, which revealed that a more diverse range of support techniques was applied to a broader range of contexts over time.

The use of agree/disagree markers exemplifies this point, as the participants gradually diversified their repertoires for doing so. At Time 1, *Me too* was the dominant method of agreement, which was surpassed by *I think so too* at Time 2, probably influenced by the latter's inclusion as a formulaic phrase in Week 4, prior to the second recording. By Time 3, however, additional forms had appeared such that there was no dominant form of agreement or disagreement (see Excerpt 27), reflecting the expanded range of techniques that is a key component of interactional competence development (Pekarek Doehler, 2019).

Another index of development is that expressions of disagreement became more nuanced and mitigated over time. Mitigated disagreement techniques (e.g., *I see your* *point, but…*) were also taught in Week 4 of the course, although they were not always used accurately at the Time 2 recording point (i.e., the mid-point of the course), as shown in Excerpt 24. Their use was both more widespread and accurate by Time 3 (see Excerpts 25 and 28), again indicating that learners needed time to practice and fine-tune the use of formulaic phrases and techniques taught in class. Incorporating elaborate forms of disagreement mirrors both L1 discourse (Pomerantz, 1984) and the behavior of more advanced L2 learners (Pekarek Doehler & Pochon-Berger, 2011). The fact that direct and unmitigated disagreement markers (e.g., *No*) occurred more at Time 1 (see Excerpt 23) corroborates Bardovi-Harlig and Salsbury's (2004) finding that these forms are more common at earlier stages of learning. Furthermore, the fact that some mitigated forms appeared following their introduction in class supports the authors' suggestion that such development can be enhanced by explicit teaching.

The range of meaning negotiation techniques, which incorporate repair and confirmations of comprehension, also expanded over time. At Time 1 (i.e., the beginning of the course), the objects of repair were overwhelmingly individual lexical items (see Excerpt 29), although by Time 3 they had broadened to include the collaborative completion of whole clauses (see Excerpts 37 to 40). Sert (2019) argued that these kinds of collaborative turn completions indicate interactional competence development because they demonstrate more attentive and active listenership than word searches. Specifically, they involve format tying at the morpho-syntactic level, which is inherently more complex than a lexical offer. They also require greater alignment and understanding of the wider point the speaker is trying to express than offers of individual lexical items. There were no instances of such completions at either Time 1 or Time 2 in the current study, which suggests a clear indication of development in this aspect of interactional competence.

An additional method of repair was observed from Time 2 onwards, as several participants rephrased their own questions when faced with apparent non-comprehension (see Excerpt 34). At Time 3, rephrasing was also used to summarize and reformulate a previous speaker's opinion, either as a precursor to changing the topic or to offering a contrasting opinion (see Excerpts 35 and 36), illustrating how the same technique was applied to a broader range of contexts as the course progressed. This finding is probably influenced by the fact that summarizing and reformulating techniques had been taught in Week 12, prior to the third recording, illustrating the reflexive relationship between pedagogy and interaction (Seedhouse, 2004). These results support Hellermann's (2011) finding that novice learners had diversified their methods of repair, and expanded the objects of repair beyond lexical items, by the end of a year-long study. Another similarity with the current study is that Hellermann's participants attempted peer repair at the earliest available opportunity (i.e., in the immediately subsequent turn) even in the initial stages of learning, suggesting the transfer of an underlying L1—or even pre-linguistic—competence.

There was also a gradual reduction in L1 use during repair sequences such that by Time 3 it was almost completely absent (see Excerpt 32), despite its obvious convenience for the participants. In the absence of sufficient L2 knowledge, the fact that gesture was retained as a method of conveying meaning, whereas L1 was not, implies a tacit recognition of the universal comprehensibility of iconic gesture and mime (Levinson, 2006). This pattern of L1 avoidance suggests a further adaptation to the interactional environment, even though during the course L1 use was not proscribed for word search exchanges. This finding supports Siegel and Seedhouse's (2024) observation that Japanese university students reduced code-switching frequency over time when attempting to resolve word search repair. Such strategies were retained to a greater extent than in the current study, however, perhaps because Siegel and Seedhouse's research was based on self-recordings made in a student dormitory, rather than in a classroom with an instructor present.

To sum up, the participants improved several aspects of their interactional competence during this one-semester course. They engaged in collaborative behavior from the earliest stages, implying that their relatively low proficiency level was not a barrier to this kind of interaction. However, qualitative analysis revealed that the techniques for doing so diversified over time, as exemplified by the range of methods for marking (dis)agreement and negotiating meaning that gradually emerged, some of which were undoubtedly influenced by their proximity to lessons in which those features had been presented and practiced.

Another indicator of development was the progression towards fewer but more complex topic extensions, as dictated by the interactional context. Moreover, the fact that they emerged concurrently with the kinds of collaborative sequences shown in Excerpts 37 to 40 illustrates that these two aspects of development are not mutually exclusive. Commenting on a topic in depth and providing listener support represent different facets of competence and can occur during the same interaction, as both contribute to the co-construction of a discussion within an academic context.

In addition, the increased use of turn and topic management features, which became smoother over time, reflected a greater willingness to take the initiative. A reduced reliance on gesture and L1 was also evident during such transitions, as well as during negotiation sequences, implying that the participants had become more confident in their ability to successfully manage interaction without recourse to their dominant language. Whether this suggestion matches the individual difference data is explored in the next section.

Research Question 3: The Extent of Speaking Self-Efficacy and Anxiety Change

As well as the performance variables of fluency and interactional competence, it is also important to investigate the affective dimension in order to understand the learner perspective. Accordingly, research question 3 concerned the extent to which the participants' speaking self-efficacy and anxiety changed during a one-semester course of spoken interaction. Regarding longitudinal studies, speaking self-efficacy has been observed to significantly increase during a year-long course of communicative interaction (Leeming, 2017) and, conversely, to decrease when opportunities for practice are limited (Busse & Walter, 2013). On the other hand, Pan & Zhang (2023) found that, following a drop in the first week, speaking anxiety remained relatively stable during a one-semester communicative course. In the current study, questionnaire data were gathered prior to all three recordings. The results were then converted to logits, produced by Rasch analysis, to investigate any changes in speaking self-efficacy and anxiety over time.

Summary of the Speaking Self-Efficacy and Anxiety Results

At the beginning of the course, the participants' self-reported scores were slightly higher for speaking self-efficacy than anxiety. Over time, a small but significant increase in speaking self-efficacy occurred between Time 1 and Time 3 (d = 0.68) and between Times 1 and 2 (d = 0.55), but not between Times 2 and 3 (d = 0.13). Therefore, the gains in self-efficacy took place overwhelmingly in the first half of the course. On the other hand, no significant changes were observed regarding speaking anxiety, despite a downward trend that produced an effect size of d = 0.37 between Time 1 and Time 3. Between Times 1 and 2 the effect size was d = 0.23, and between Times 2 and 3 it was even smaller at d = 0.14. It can therefore be concluded that speaking anxiety remained relatively stable throughout the course.

Interpretation of the Speaking Self-Efficacy and Anxiety Results

Regarding speaking self-efficacy, the closest contextual similarity to the current study is with that of Leeming (2017), who also investigated its development among freshman Japanese university students. In both cases, the participants had limited previous experience of English oral communication and were placed in low-intermediate classes. Leeming's study was also based on communicative speaking classes, although over a whole academic year rather than one semester, and his self-efficacy questionnaire was adapted for use in this study (see Appendix K). Leeming's instrument was administered eight times throughout the year, constituting a repeated-measures design, as opposed to three times in the current study; therefore, different statistical analyses were conducted. However, in both cases results were based on logit measures, facilitating a comparison of the descriptive statistics. Leeming's participants increased from a low starting point of -2.75 logits to -0.07 by the end of the year, following a relatively linear progression. The participants in the current study began from a higher starting point of 1.29 logits, which had increased to 2.70 by Time 3. Clear gains were therefore observed across both studies; moreover, both studies saw the biggest increase between the first two measurement points (1.14 logits in the current study; 1.26 in Leeming's study).

Taken together, these findings imply that speaking self-efficacy in EFL contexts such as these is likely to improve especially in the early part of the course, perhaps reflecting the unfamiliarity of this kind of communicative environment prior to entering university in Japan (Koizumi et al., 2022). In other words, learners are unlikely to have a strong belief in their ability to do activities with which they have had little or no previous experience. The fact that self-efficacy stabilized in the current study but continued to develop in Leeming's probably reflects the different starting points, as the participants in the current study had less room to increase after Time 2. Why the starting points should be so different, despite an apparently similar level of proficiency at the outset, is of potential interest. One possible reason is that the students in Leeming's study were all science majors, whereas the current study was conducted in a liberal arts department in which English forms a major part of the curriculum. As such, it is probably more likely to attract students who are receptive to the idea of studying English. It is beyond the scope of this study to examine differences in affect between students of different university departments, although that could represent a promising area of future research.

Regarding speaking anxiety, the results in the current study were broadly similar to that of Pan and Zhang's (2023) longitudinal study, which was also based on a single semester in a communicative, university EFL environment, albeit with upper-intermediate learners in China rather than lower-intermediate learners in Japan. In both cases, anxiety remained stable following a relatively small initial drop. Pan and Zhang's questionnaire was distributed after every lesson (42 times in total), with an average fall from about 3.50 to 3.00 on the 5-point rating scale occurring during the first week of the course, after which there was very little fluctuation. The questionnaire in the current study was only distributed three times, with a larger change occurring between Times 1 and 2 (0.40 logits) than between Times 2 and 3 (0.23 logits), although no changes reached statistical significance. The lack of inferential statistics or effect sizes in Pan and Zhang's study mean it is difficult to compare the results of the two studies directly, although they both appear to support the idea that speaking anxiety is unlikely to change significantly during one EFL semester. Furthermore, the small reductions that did occur in the early stages could have resulted from overcoming the stress of an unfamiliar environment, including a new teacher and classmates.

Looking at the relationship between these two variables, in the current study a generally small to medium-sized negative correlation was observed between speaking self-efficacy and anxiety at each time point (r = -.14., -.47, and -.34, respectively), although statistical significance was reached only at Time 2. This relationship is less robust than that established in Zhou et al.'s (2023) meta-analysis (r = -.70), although the latter was not limited to speaking studies. It nevertheless supports the idea that learners with higher levels of self-efficacy tend to exhibit lower levels of anxiety (Pajares & Schunk, 2001), as well as the idea—expressed in social cognitive theory—that anxiety is stimulated by a weakened sense of self-efficacy (Bandura, 1997).

End of Course Questionnaire Responses

The student participants were given an end of course questionnaire (see Appendix Q), which was completed anonymously and on a voluntary basis, in order to provide an emic perspective on the course. The following comments have been selected to supplement the affective data analyzed above. All questionnaire comments in this chapter have been translated from the original Japanese. The most common theme that emerged from the responses was the perception of an improved ability to express opinions in English, with comments such as “I became able to say my opinion in English” occurring frequently in answer to the question “In what way(s) do you feel you improved the most during this course?” These comments imply an increase in self-efficacy as it relates to this course, as giving opinions is an integral aspect of a discussion. Some participants also mentioned perceived improvements in other aspects of their speaking performance. For example, Excerpt Q1 refers to specific skills that the participant feels they learned, whereas Excerpt Q2 mentions an increase in fluency.

Excerpt Q1

> Not only could I convey my own thoughts, but I became able to speak while doing things like rephrasing and agreeing with the other person.

Excerpt Q2

> I’ve become able to speak English much faster than before. In particular, I feel that English words come out much easier than before.

Whereas the above excerpts are unambiguously positive, the following two excerpts imply that, for these participants, self-efficacy increased from a relatively low starting point, given the references to initially silent behavior.

Excerpt Q3

> I got a little more confident about speaking English and I'm less likely to stay silent when speaking English.

Excerpt Q4

> At first, I didn't have my own ideas, I couldn't speak English, and I was often silent. But now I've got the attitude that anything's okay so just speak.

As well as increased self-efficacy, some responses also implied a reduction in anxiety, despite the fact that, according to the questionnaire data, it remained relatively stable throughout the course. Both Excerpts Q5 and Q6 imply reduced anxiety (i.e., feeling "nervous" or "worried"), either combined with a perception of greater fluency, in the former case, or the ability to participate more actively, in the latter.

Excerpt Q5

> When I spoke to others in English, I became able to speak more smoothly and without feeling nervous.

Excerpt Q6

> I became able to say my opinion in English without hesitation or feeling worried. I became able to speak actively.

All student participants chose to complete the questionnaire and there were no negative responses, despite it being conducted anonymously, which implies a widespread and generally positive attitude towards the course. The vast majority of responses referred to things that the participants felt able to do or had improved, which supports the quantitative data regarding increased speaking self-efficacy. Moreover, as Excerpts Q3 and Q4 show, even students who started from a lower base appeared to show increases in self-efficacy. In addition, several responses mentioned feeling less nervous or worried, which implies that at least some participants felt a reduction in speaking anxiety.

Research Question 4: The Extent to Which Speaking Fluency Is Related to Interactional Competence

Research question 4 concerned the extent to which speaking fluency is related to interactional competence. Although some studies have found moderate correlations between general speaking proficiency, as measured on the TOEFL iBT test, and ratings of interactional competence (Ockey et al., 2015; Roever & Ikeda, 2021), little is known about the relationship between quantitative measures of interactional competence and speaking fluency. Correlation analysis was used to address this question, although the eight categories of interactional features were aggregated into only two domains (i.e., topic development and support moves) for this purpose. These two domains were correlated with the three fluency variables: utterance fluency (i.e., pruned speech rate), fluent syllables per minute produced in the group discussions, and perceived fluency.

Summary of the Results for the Relationship Between Speaking Fluency and Interactional Competence

As mentioned previously, the three fluency variables were all significantly and positively correlated at each time point, with large effect sizes in almost every case. This was also true of the two domains of interactional competence features, indicating that the participants who contributed the most topic development moves also tended to contribute the most support moves, and vice versa. The only clear relationships between the two constructs, however, were with fluent syllables per minute produced in the group discussions and the number of both topic development and support moves. These relationships were statistically significant, and with relatively large effect sizes, at Times 1 and 2 (r = .47 and .60, respectively, at Time 1; r = .63 and .58, respectively, at Time 2). However, they were non-significant, with medium or small effect sizes, at Time 3 (r = .39 and .22, respectively). There were no consistent relationships between the interactional competence measures and either utterance fluency (i.e., pruned speech rate) or perceived fluency, although all were at their weakest at Time 3 (see Tables 53 to 55 in Chapter 5 for the correlation matrices).

Post-hoc analysis revealed significant relationships between words per topic extension and all three fluency variables at Time 3. Specifically, with syllables per minute the relationship was r = .67, with utterance fluency it was r = .58, and with perceived fluency it was r = .45. The strongest relationship, however, was with communicative adequacy (r = .72), implying that the longer and more detailed contributions at the end of the course were detected and positively rated.

Interpretation of the Results for the Relationship Between Speaking Fluency and Interactional Competence

It is perhaps unsurprising that the more interactional moves the participants contributed, the more syllables they generally produced during the group discussions. However, the fact that the relationships between the interactional competence measures and all of the fluency variables were at their weakest at Time 3 is of potential interest. Regarding support moves, the effect sizes at this time point—all non-significant—were r = .22, .17, and .05 for fluent syllables per minute, utterance fluency, and perceived fluency, respectively. Taken together, these findings suggest that by Time 3 less fluent participants had almost caught up with their more fluent peers in terms of their ability to offer listener support (i.e., provide backchannels, agree/disagree, and engage in the negotiation of meaning), implying that lack of fluency is not necessarily a barrier to the development of effective interaction. However, these findings should be interpreted with caution given that the four categories of support moves were aggregated into one domain for the purposes of statistical analysis.

A broadly similar pattern can be observed regarding topic development moves. The effect sizes at Time 3—again all non-significant—were r = .39, .16, and -.03 for fluent syllables per minute, utterance fluency, and perceived fluency, respectively. At first glance, these results might suggest that less fluent participants had also caught up in terms of their ability to develop the topic under discussion, for example by providing topic extension statements or questions. However, this picture is complicated by the fact that the number of topic extensions significantly reduced over time, as the participants began to produce longer and more detailed comments.

Again, it is unsurprising that longer topic extensions correlated with overall syllables produced, but the fact that they also correlated with utterance and perceived fluency at Time 3 indicates that this trend was related not only to the development of interactional competence but also with other aspects of speaking proficiency. For example, utterance fluency development is associated with automaticity and parallel processing (Kormos, 2006); therefore, increased utterance fluency, as measured on the monologue activities, could enable learners to produce longer and more complex topic extensions during the group discussions by freeing up cognitive processing capacity. On the other hand, if more fluent speakers become more adept at producing such comments over time, less fluent speakers are at a disadvantage in terms of being able to adapt to this aspect of the task and therefore the interactional context.

To sum up the relationship between fluency and interactional competence, these results imply that lack of fluency was not necessarily a barrier to interaction in the sense of offering listener support, as less fluent participants produced almost as many support moves as more fluent participants at Time 3. However, fluency did appear to present a challenge in terms of contributing more detailed and complex topic extensions. This claim is supported by the strong positive relationships between words per topic extension and all aspects of fluency at the end of the course, as it implies that more fluent speakers were able to develop and expand such comments to a greater extent than their less fluent peers.

Research Question 5: The Extent to Which Speaking Fluency Is Related to Speaking Self-Efficacy, Speaking Anxiety, and Extraversion

Research question 5 concerned the extent to which L2 speaking fluency is related to speaking self-efficacy, speaking anxiety, and extraversion. Previous research has not addressed the relationships between all of these individual difference variables and all the aspects of fluency included in this study. However, of those that are most directly related to this study, a negative relationship has been observed between speaking anxiety and several aspects of utterance fluency (Bielak, 2022; Pérez Castillejo, 2019). On the other hand, a positive relationship has been observed between speaking self-efficacy and the number of words produced in a group discussion (Leeming et al., 2024), which is comparable to the syllables produced measure used in this study.

The other individual difference variable examined in this study was extraversion, which has been found to correlate positively with speaking self-efficacy (Leeming, 2017) and negatively with speaking anxiety (Pan & Zhang, 2023). Regarding its relationship with fluency, extraversion has been observed to correlate positively with the composite measure of speech rate and negatively with rate of filled pauses, although the latter result was produced only in formal situations (Dewaele & Furnham, 2000). Furthermore, Ockey (2011) found that assertiveness—defined as one aspect of extraversion—was a significant explanatory variable of perceived fluency, albeit with a small effect size. Correlation analysis was used to address these relationships in the current study.

Summary of the Results for the Relationships Between Speaking Fluency and Speaking Self-Efficacy, Speaking Anxiety, and Extraversion

Fluency and extraversion had the clearest relationship in this study, which was positive and consistent at each time point. Specifically, the relationship with fluent syllables produced in the group discussions was strong and statistically significant (r = .58., .68, and .63 at Times 1, 2, and 3, respectively); with utterance fluency it was medium (r = .32., .57, and .44), although significant only at Time 2; and with perceived fluency it was small and non-significant (r = .33, .29, and .36). Speaking self-efficacy had a significant, medium-sized relationship with syllables produced (r = .46), and a small, non-significant relationship with utterance fluency (r = .24), but only at Times 2 in both cases. In contrast, a medium to large negative relationship was observed between speaking anxiety and syllables produced, and a small to medium negative relationship was observed with utterance fluency at Times 1 and 2 (r = -.33 and -.66, respectively, for the former; r = -.24 and -.49 for the latter), although significant only at Time 2 in both cases. No clear associations were found between either of these two variables and perceived fluency at any time point. See Tables 56 to 58 in Chapter 5 for the correlation matrices.

Interpretation of the Results for the Relationships Between Speaking Fluency and Speaking Self-Efficacy, Speaking Anxiety, and Extraversion

The results outlined above indicate that more extraverted participants performed more fluently throughout the course, especially in terms of syllables produced in the group discussions, and to a lesser extent utterance and perceived fluency. In addition, the fact that the extraversion findings were the most stable over time supports the idea that, as a dispositional personality trait, its effects are more predictable than contextualized characteristics such as self-efficacy and anxiety. On the other hand, as context-dependent variables are more liable to change during a course of study, the effects they have on outcomes such as fluency and quantity of output is also more likely to vary.

Of the limited number of other studies that can be directly compared with these results, Dewaele and Furnham's (2000) findings regarding utterance fluency were broadly similar. They found a positive, significant, and fairly strong correlation between extraversion and speech rate, both for assessed and unassessed conversations (r = .51 and .55, respectively). In the current study, which was based on unassessed discussions, the effect was moderate (r = .32., .57, and .44 at Times 1, 2, and 3, respectively), although significant only at Time 2. The current study, therefore, offers some support to the idea that extraversion is positively related to utterance fluency. Dewaele and Furnham (2000) argued that this association results from an over-aroused autonomous nervous system on the part of introverts, which can overload working memory and affect parallel processing, leading to less fluent speech production.

Regarding speaking self-efficacy, the only significant relationship in this study was with fluent syllables produced in the group discussions at Time 2. This result offers some support for Leeming et al.'s (2024) finding that speaking self-efficacy was positively related to total amount of words produced in a group discussion, as both are general measures of speech rate within a communicative context. Leeming et al. found a significant, medium-sized correlation of r = .41, whereas in the current study the correlation was r = .15 and .46 at Times 1 and 2, respectively. However, the fact that this relationship had completely disappeared by Time 3—as had the weaker relationship with utterance fluency—implies that lower speaking self-efficacy was not an impediment to fluent performance at the end of the course.

A broadly similar, but inverse, pattern was observed regarding the relationship between speaking fluency and anxiety. Regarding utterance fluency, this result supports Pérez Castillejo (2019) and Bielak (2022), both of whom found that language anxiety correlated negatively with several individual speed and breakdown fluency measures (as opposed to the composite measure of pruned speech rate used in this study). However, this relationship had again disappeared by Time 3 in the current study, as had the relationship with fluent syllables per minute in the group discussions. In other words, lower speaking anxiety, as with higher speaking self-efficacy, was associated with more fluent performance in the first half of the course but not at the end. The similarity of these two patterns suggests that students with lower positive affect were able to overcome this disadvantage in the second half of the course.

Looking at some individual cases illustrates this point. For example, Aoi reported the second lowest speaking self-efficacy of all participants at Time 3, as well as the fourth highest speaking anxiety. However, she produced the most fluent syllables per minute in the group discussions (64.89) and had the highest utterance fluency (i.e., pruned speech rate) in her monologue (151.33 syllables per minute) at the same time point. Taking another example, Kiyomi's self-reported speaking self-efficacy actually fell throughout the course and was the seventh lowest at Time 3. Furthermore, she reported consistently high speaking anxiety and the second highest at Time 3. However, her utterance fluency increased from 63.33 syllables per minute at Time 1 to 102.00 at Time 3, which took her from the bottom half to the top half of participants. In addition, her fluent syllables per minute in the group discussions almost doubled from 16.33 at Time 1 to 31.00 at Time 3. These two cases emphasize the fact that low or below average affect was less of a barrier to fluent performance at the end of the course.

Research Question 6: The Extent to Which Interactional Competence Is Related to Speaking Self-Efficacy, Speaking Anxiety, and Extraversion

Research question 6 concerned the extent to which interactional competence is related to speaking self-efficacy, speaking anxiety, and extraversion. Little research has addressed any of these points directly, although some studies have investigated specific aspects of the relationship between extraversion and interactional competence. For example, Nakatsuhara (2013) found that extraverts took quantifiably more responsibility for managing turns and topics during group discussions. Also, Ockey (2011) found that assertiveness—as a facet of extraversion—was a significant explanatory variable of rated interactional skills, as well as fluency, in group discussions. Correlation analysis was used to address these relationships in the current study, using the two domains of aggregated interactional moves.

Summary of the Results for the Relationships Between Interactional Competence and Speaking Self-Efficacy, Speaking Anxiety, and Extraversion

As with fluency, the clearest and most consistent relationship between interactional competence and the individual difference variables was with extraversion. Specifically, the relationship with both topic development and support moves was moderately positive at each time point (r = .36, .55, and .44 at Times 1, 2, and 3, respectively, for the former; r = .38, .58, and .3 at for the latter), although only reached statistical significance at Time 2 in both cases. Regarding speaking self-efficacy, a small to moderate positive, although non-significant, relationship was found with topic development and support moves at Time 1 (r = .25 and .40, respectively). The relationship then disappeared in the former case, but was maintained in the latter, albeit in weaker form (r = .24 and .31 at Times 2 and 3, respectively). Regarding speaking anxiety, a generally weak and non-significant negative relationship with topic development moves existed throughout (r = -.14, -.32, and -.23 at Times 1, 2, and 3, respectively). A moderately negative relationship was observed with support moves at Times 2 and 3 (r = -.41 and -.46, respectively), which was significant only in the latter case (see Tables 59 to 61 in Chapter 5 for the correlation matrices). Post-hoc analysis revealed inconsistent relationships between words per topic extension and both speaking self-efficacy and anxiety, although none of these correlations were statistically significant.

Interpretation of the Results for the Relationships Between Interactional Competence and Speaking Self-Efficacy, Speaking Anxiety, and Extraversion

These results support previous findings that extraversion is positively related to aspects of interactional competence in a group discussion context (Nakatsuhara, 2013). This finding might reflect the fact that assertive participants are more likely to take the lead during interaction (Ockey, 2011). It might also reflect the demands on working memory that simultaneously activating speaker and listener roles entails, as introverts are known to have weaker working memory skills than extraverts (Lieberman, 2000).

Turning to the other individual difference variables, both the positive relationship between speaking self-efficacy and support moves, and the generally negative relationship between speaking anxiety and support moves, imply that positive affect plays at least some role in facilitating engagement with interlocutors and other speakers' contributions. Regarding topic development moves, however, the picture is once again complicated by the reduction in the number of topic extensions that took place over time. As discussed previously, fluent words per topic extension is arguably a better measure of competence in this context, as it implies engaging with the topic in depth; therefore, this measure was additionally correlated with the individual difference variables (see Tables 69 to 71 in Post-Hoc Analyses).

What appears to have happened is that participants higher in self-efficacy made lots of short contributions at Time 1, resulting in a negative correlation with words per topic extension. However, these participants adapted to the discussion task more quickly, producing longer contributions at Time 2, thereby reversing the polarity of the relationship. Finally, participants lower in self-efficacy had caught up by Time 3 such that the polarity was reversed once more. This interpretation is supported by the correlations between number of topic development moves—which consists mainly of topic extension statements—and speaking self-efficacy (see Tables 59 to 61), as this relationship went from positive at Time 1 (r = .25) to virtually non-existent at Times 2 and 3 (r = .09 and -.02, respectively). In other words, self-efficacy shifted from an association with many short contributions at the start of the course to one with fewer but longer contributions by the mid-point, which then disappeared at the end of the course as lower self-efficacy was no longer an impediment to producing longer topic extensions. Similarly, speaking anxiety was no longer negatively associated with words per topic extension at Time 3 (r = .20), having reversed polarity from r = -.16 and -.34 at Times 1 and 2. These findings recall the relationships between speaking self-efficacy, anxiety, and fluency—discussed in the previous section—whereby students with lower positive affect had overcome this disadvantage by the end of the course.

Again, individual cases illustrate the point. For example, Konomi consistently reported low speaking self-efficacy, which was the lowest of all participants at Time 3. However, at the same time point she contributed the fourth highest number of support moves (21) and increased the length of her topic extension statements to an average of 36.33 words, from 13.80 at Time 1 and 18.50 at Time 2. Another example is Yukino, who reported the third lowest speaking self-efficacy as well as the third highest speaking anxiety at Time 3. However, she produced the fifth highest words per topic extension (41.17) at that time point, which increased from 16.78 at Time 2. The pedagogical implications of these findings are addressed at the end of this chapter.

Theoretical Implications of the Results

These findings have several theoretical implications for researchers interested in the development of second language speaking fluency and interactional competence. The first regards the quantification of speaking fluency development. The results of this study add to the growing body of literature that suggests EFL learners in instructed settings can make observable gains in fluency during just a single semester of study in a communicative environment, even if they have limited previous experience of using the target language productively (Doe, 2025; Hanzawa, 2021). Although the monologue-based gains in utterance fluency were small, they followed developmental patterns identified by other researchers, as initial gains were based mainly on reduced silent pause duration whereas later gains were accumulated from several different measures. Among utterance fluency measures, De Jong et al. (2013) found that gains in silent pause reduction were the least related to underlying cognitive fluency and argued that they could therefore be due to non-linguistic factors such as individual differences. The fact that, in this study, the timing of these initial fluency gains coincided with a reported increase in speaking self-efficacy seems to support this theory. On the other hand, accumulated gains from several aspects of the speech production process, as happened in the second half of the course, indicate better overall integration of these features, and therefore more efficient cognitive processing (Segalowitz, 2010).

The fact that these monologue-based fluency gains were corroborated by measures taken from the interactive context further supports the idea that fluency had developed during the course. Regarding group-based measures, there was an increase in fluent syllables per minute that was detected by raters at the end of the course. These results were further supported by a reduction in the number and mean length of gaps between turns. To sum up, although the extent to which monologue-based and interactional fluency represent the same construct is debatable (Tavakoli, 2016), the fact that all these results point in the same direction suggests that measures taken from interaction can assist researchers in triangulating more traditional, monologue-based measures of fluency.

Although there is a stronger tradition of measuring L2 speaking fluency in SLA research, this study also supports the idea that some aspects of interactional competence can be quantified. However, due to the highly context-dependent nature of interactional competence researchers should tailor their coding schemes to the specific situation under investigation, as features that are appropriate in one context might not be so in another (Schegloff, 1993). Moreover, due to the contingencies of interaction, opportunities to use potentially appropriate features (e.g., repair) might not always arise or seem necessary to participants. Van Batenburg et al. (2018) circumvented this issue by tightly controlling interlocutor contributions (e.g., including unfamiliar words in examiner scripts) to elicit specific communicative strategies, and by doing so the authors were able to reliably measure interactional ability. However, in the case of freer speaking activities—such as group discussions—there is no guarantee that certain interactional features will be manifested, and therefore that individuals will be able to demonstrate the full range of their competence. Nevertheless, given a sufficient amount of data, quantification can identify patterns of interaction that distinguish between and more less competent participants (Galaczi, 2008).

In the current study, significant change in the number of interactional moves was limited to an increase in turn/topic management moves and a decrease in topic extensions. The latter was followed up by analysis that revealed an increase in the mean number of words per topic extension statement, demonstrating how participants had adapted to the demands of the discussion activity. The fact that no significant change occurred regarding any of the support features emphasizes the need for qualitative to complement quantitative analysis, as development in this case was revealed only by careful examination of the content produced and the phrases utilized. Galaczi (2014), on the other hand, was able to identify differences among proficiency levels in terms of both topic development and support moves, albeit in a different speaking context. This disparity raises the possibility that quantifying interactional competence is better suited to distinguishing between levels than highlighting longitudinal development within them. It is also possible that quantifying IC moves might be better suited to highlighting development over longer terms than one semester. Nevertheless, the patterns that emerged from measuring individual performance in this study—within changing group configurations—suggest that the ability to interact can be transferred to novel instances, consisting of different participants, despite the contingencies of multi-party interaction. Therefore, the insights gained by quantification in the current study support Fulcher's (2010) contention that it is primarily the resources that individuals bring to interaction—albeit mediated by context and its myriad external variables—that constitute interactional competence.

Pedagogical Implications of the Results

The results of this study also have several pedagogical implications. First, the triangulated gains in speaking fluency support the idea that extensive practice is necessary to proceduralize—and eventually automatize—linguistic knowledge, as proposed in skill acquisition theory (Anderson, 1992; DeKeyser, 2007). Students in this discussion-based course were given many opportunities, either in pairs or groups, to talk about a range of topics related to university life and society in general (see Table 1 in Chapter 3 for the course schedule). Every lesson also included a repeated monologue activity to further encourage fluency development (Nation, 1989). The class met twice a week, with the second lesson a textbook-based speaking lesson that also included pair, group, and fluency-building activities. This course, therefore, clearly fulfills the criteria for a communicative approach to language teaching, in which meaning-focused communication is prioritized, as recommended in several pedagogical frameworks that aim to develop L2 fluency (e.g., Gatbonton & Segalowitz, 2005; Nation, 2007). The fact that it was successful in developing participants' speaking fluency in one semester, and that this development extended beyond the initial phase of reduced silent pausing time, demonstrates the benefits of such an approach. Therefore, the reported lack of teaching resources, syllabi, or textbook activities that reflect these practical implications (Rossiter et al., 2010; Tavakoli, 2023) is a situation that needs to be addressed, especially in EFL contexts in which opportunities for practice outside of the classroom are often limited.

A second implication concerns the direct teaching of interactional skills and strategies as a way of improving interactional competence. The excerpts taken from the group discussions show how participants increased their range of available resources for accomplishing these skills—such as managing turns and topics, indicating (dis)agreement, and negotiating meaning—more smoothly and without relying on the L1. This pedagogical approach therefore fostered the flexibility necessary to deal with L2 interaction collaboratively and efficiently, which was also reflected in the reduced number and length of gaps between turns. Collaborative behavior, which is at the heart of interactional competence, has been found to produce higher scores on paired speaking tests (Artunc & Hart, 2019; Galaczi, 2008). The results of this study indicate that such behavior is also essential to the smooth operation of unassessed group discussions, and that it can be directly taught via micro-skills and strategies (Galaczi, 2008; Richards, 1990).

Furthermore, the fact that the relationships between utterance fluency, from which underlying fluency can be inferred, and number of topic development and support moves had virtually disappeared by the end of the course implies that lower speaking fluency was not an impediment to their use over time. In other words, less fluent learners had caught up with their more fluent peers in terms of the amount of such features they used.

This finding further implies that these skills, and by extension interactional competence, can and should be taught because their use did not appear to depend on participants' levels of fluency.

The pedagogical value of these skills and strategies is reinforced by ascertaining the emic perspective of the participants. Prior to the recordings, instructions were given for participants to discuss the topic in the way they felt was best, and that it was not necessary to use what they had learned in class (see Appendix J for the written instructions). Although desirability effects cannot be ruled out, the intention was for the learners to use these techniques and phrases only if they felt they were beneficial. The exemplars provided in the Chapter 5 excerpts give an indication of which skills the participants valued. Another indication is provided by the end of course questionnaire. Figure 17 (below) shows the number of times an interactional feature was mentioned in answer to the question "Which discussion skills, phrases, or questions did you find the most useful in this course?" For example, mentioning the importance of giving or supporting an opinion was counted as an instance of TES (i.e., topic extension statement). Respondents could mention more than one category. If responses indicated specific phrases or questions that were taught, such as *What do you think?* and *I partly agree with you*, they have been included under the relevant category.

The two categories with the most mentions were agree/disagree markers and confirmations of comprehension (11 and 10, respectively), which might be counter-intuitive as their use did not change quantifiably during the course. There were qualitative improvements in their use, however, which seem to have been noticed by the participants based on these responses. It is also notable that topic extension statements had the second lowest number of mentions, given that there was considerable development in the ability to produce detailed and complex comments. As stated previously, many students mentioned an improved ability to express opinions in English in answer to the earlier question “In what way(s) do you feel you improved the most during this course?” so they might have been reluctant to give a similar response to the later question. Another possibility is that most participants did not equate giving their opinion with performing a specific discussion skill, but more generally as just speaking English in this context (i.e., carrying out the task).

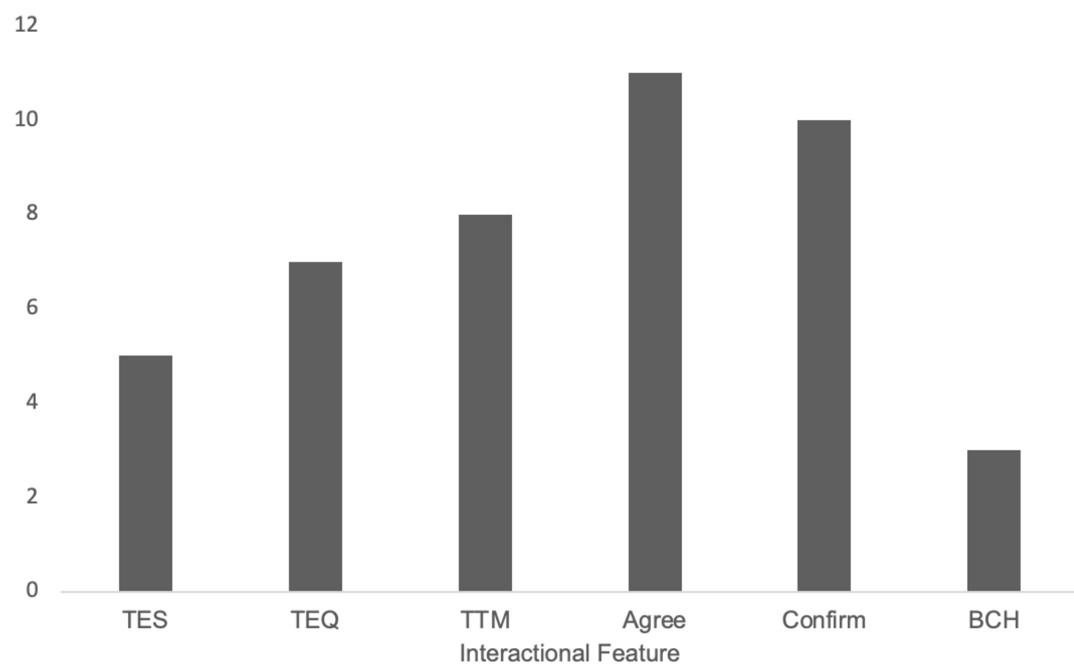

Figure 17. Participant Mentions of Most Useful Interactional Features
Note. TES = topic extension statement, TEQ = topic extension question, TTM = turn/topic management, Agree = agree/disagree marker, Confirm = confirmation of comprehension, BCH = backchannel/reaction.

Some of the comments are reproduced in full below to provide further insight into the participants’ perspectives regarding the pedagogical materials and skills that were taught. Excerpt Q7 refers to the practical benefits of learning formulaic phrases, whereas Excerpts Q8 and Q9 also point out that they can provide an initial impetus to begin speaking.

Excerpt Q7

> I learned a lot of expressions that can actually be used in conversation and this developed my knowledge.

Excerpt Q8

> I think it's very useful to learn fixed sentences and phrases because they are also an opportunity to start speaking.

Excerpt Q9

> After saying whether you agree or disagree with someone's opinion you can start to give your own opinion.

As well as helping learners express their own ideas, some of the skills were aimed at developing listener support. The following two excerpts highlight the importance of such support for sustaining the interaction, by encouraging others to participate (Excerpt Q10) and asking questions (Excerpt Q11).

Excerpt Q10

I felt that the ability to encourage others to speak in the discussion was important. In order to have a smooth discussion it's an important skill.

Excerpt Q11

I realized the importance of discussion skills that can continue the conversation. I think the conversation will be lively if you can understand and summarize your partner's opinions and ask related questions.

Despite the recent growth in interactional competence research, there have been few attempts to connect such research with pedagogy (Salaberry & Kunitz, 2019). However, it is clear from these responses that the students valued the techniques they were taught during the course. This emic perspective complements the etic perspective of the researcher; moreover, the fact that both perspectives are positively aligned adds to the pedagogical value of this direct approach to teaching such skills. Furthermore, the fact that giving opinions, agreeing/disagreeing, and turn taking were all highlighted as aspects of interactional competence development in this study, and were mentioned positively in questionnaire responses, justifies their inclusion in the early weeks of the course (see Table 1 in Chapter 3).

The final pedagogical implication is that speaking self-efficacy can increase as a result of this kind of communicative teaching, even if anxiety remains relatively stable. Goetze & Driver's (2022) meta-analysis established a positive link between L2 self-efficacy and achievement, in which the strongest relationship was with speaking self-efficacy. It is therefore beneficial to conduct activities that develop positive affect. Mills (2014) stated that speaking self-efficacy can be fostered by providing learners with regular opportunities to exchange information and opinions with their peers in a comfortable environment. The course upon which this study was based provided many opportunities for the participants to exchange opinions, resulting in a significant increase in reported self-efficacy. These opportunities facilitated the practice effect mentioned earlier, which can not only lead to increased fluency but also self-efficacy as successful experiences promote feelings of mastery among participants (Bandura, 1997).

However, the picture was more complicated than this finding suggests. Whether looked at from a fluency perspective (i.e., total syllables produced), or an interactional perspective (i.e., words per topic extension) results showed that speaking self-efficacy became less associated with participation over time. In other words, students with lower self-efficacy had caught up with those higher in self-efficacy by the end of the course in this regard. As mentioned above, the benefits of self-efficacy regarding achievement are clear; however, if even students who have relatively low speaking self-efficacy, or high speaking anxiety, can overcome these disadvantages and participate actively in a communicative environment, such a course has even greater benefits. The following two comments, taken from the end of course questionnaire, shed some light on how this works in practice.

Excerpt Q12

> I feel like the hurdle of speaking English has been lowered. I've come to think that it's okay to speak even if I'm not good at it.

Excerpt Q13

> It's important to say your opinion using the knowledge and vocabulary that you have. Even when I couldn't, I could ask my friends around me so it became easier to say my opinion in English.

These comments suggest that even students who felt unable to express themselves well in English were still able to play an active part in the class. They therefore support the quantitative questionnaire results, which showed that lower self-efficacy was not an impediment to participation at the end of the course, especially regarding the number and length of topic extensions. The resulting practice effect could also explain the disappearance of the relationship between speaking self-efficacy and utterance fluency by the end of the course. Furthermore, the lack of a negative relationship between speaking anxiety and any of the fluency variables, or words per topic extension, at the end of the course suggests that it does not prevent students from developing their fluency or their ability to make extended contributions to a discussion. In sum, these findings imply that educators should strive to create an environment that not only fosters self-efficacy but also enables students whose self-efficacy remains low, or whose anxiety remains high, to feel able to participate.

Summary of Discussion

In this chapter, I summarized and discussed the findings for each of the six research questions. In addition, I related these findings to previous studies and existing research in the relevant fields. Finally, theoretical and pedagogical implications of the study were addressed. Next, in Chapter 7, I briefly summarize the main findings of the study. I then address its limitations and offer suggestions for future research in the area of L2 speaking development. I conclude with some final comments.

CHAPTER 7

CONCLUSION

In this chapter, I begin by summarizing the main findings of the study. After that, I discuss some of its limitations, then offer suggestions for future research. I conclude the study with some final comments about what led me to conduct this research and what I hope it has achieved.

Summary of the Findings

This study was designed to investigate the extent to which Japanese university students' speaking fluency and interactional competence developed during a one-semester communicative course, based on the teaching of interactional strategies and techniques. Many learners express the desire to speak a language fluently, although in practice this goal entails not only speaking relatively quickly and smoothly but also interacting with others in ways that are appropriate to the communicative context. However, it remains an open question whether and how quickly learners can acquire such skills in an EFL environment, in which opportunities to use the target language are limited outside the classroom. An additional purpose of this study was to investigate any changes in speaking self-efficacy and anxiety during the course, and how these individual difference variables related to the performance variables of speaking fluency and interactional competence. To address these issues, data were gathered from 20 low-intermediate learners at three time points throughout the semester.

The first main finding, based on a variety of measures, was that speaking fluency significantly improved over the course of the semester. Regarding utterance fluency, as measured on monologue activities, gains in the first half of the course were mainly due to a reduction in pausing time, which is a sign of incipient fluency development (Hanzawa, 2021). In the second half of the semester, however, an accumulation of marginal gains across various measures implies that discrete aspects of speech production were integrating, and thus working more efficiently, which suggests the beginnings of underlying cognitive fluency development (Segalowitz, 2010). Furthermore, these utterance fluency gains were matched by gains made in the interactive context. Specifically, there was an increase in the number of syllables that the participants produced during the group discussions, as well as a reduction in the number and length of inter-turn pauses. These gains were also detected by raters, who judged the participants significantly higher in fluency, and communicative adequacy, at the end of the course. Taken together, these findings present a powerful case for the fluency-enhancing benefits of communicative interaction in an EFL environment.

The second main finding, based on a combination of quantitative and qualitative analysis, was that interactional competence also developed during the semester. Quantitative analysis revealed a significant increase in the number of turn and topic management moves over time, indicating an increase in collaborative behavior as well as a greater willingness to take the initiative in managing the interaction. In contrast, there was a significant decrease in the number of topic extension statements, including minimal extensions, with the participants making longer and more complex contributions over time, as they adapted to the demands of the task and thus the interactional context. This development favored more fluent participants, as shown by the strong positive relationship between words per topic extension and all aspects of fluency at Time 3. However, there was no such relationship between utterance fluency and the number of interactional moves—both topic development and support—at the end of the course, which implies that using these skills was not dependent on the participants' levels of fluency.

Qualitative analysis was used for a detailed examination of the micro-skills and techniques that constitute interactional competence (Galaczi & Taylor, 2018). This analysis confirmed the shift from short, often sub-clausal topic extensions to longer, multi-clausal extensions, which coincided with a shift from short, often lexical topic extension questions to longer, more precisely targeted questions. There was also a gradual progression towards the smoother use of turn and topic management features, and a diversification of techniques for marking agreement and disagreement. In addition, techniques for negotiating meaning were applied to a broader range of situations, and several instances of collaborative turn completions emerged at the final data collection point. Finally, there was a reduced reliance on L1 when negotiating meaning and transitioning between turns and topics. These findings emphasize the fact that, although quantification can reveal some aspects of development, qualitative analysis is necessary to reveal the essence of how interactional competence develops in context.

The third main finding was that speaking self-efficacy increased significantly in the first half of the course, which probably reflects the unfamiliarity of communicative environments for most pre-university learners in Japan. The form-focused manner in which English is generally taught prior to university means that learners who do well in high school have often accrued a good deal of declarative knowledge, which DeKeyser (2010) stated is a prerequisite for proceduralization through communicative practice. Therefore, when given the opportunity to use that knowledge productively such learners often realize that they have a solid foundation on which to develop their communicative skills. This realization then leads to the kinds of successful experiences that promote heightened self-efficacy (Bandura, 1997). The increased measure of speaking self-efficacy correlated with several of the performance variables (i.e., utterance fluency, number of syllables produced in the discussions, and length of topic extension) at the mid point of the course. However, the disappearance of these positive correlations at the end of the course implies that students lower in positive affect had caught up with their peers in these aspects, which implies even wider benefits of this approach.

In summary, the learners in this course made demonstrable gains in terms of speaking fluency, interactional competence, and speaking self-efficacy. Moreover, these elements appeared to work in concert, creating a positive feedback loop whereby higher self-efficacy coincided with longer topic extensions, facilitated by more efficient turn taking, which provided more practice, thus creating the conditions necessary for fluency development. Underpinning all of this was the communicative environment, in which learners were provided with plentiful opportunities to exchange ideas and opinions, and practice the skills that were taught, with the benefits of this approach ultimately extending even to participants lower in fluency and speaking self-efficacy.

Limitations of the Study

This study had several limitations that should be taken into account when interpreting the results. The first limitation concerns the number of core participants, as an *N*-size of 20 raises issues of statistical power for the analyses conducted. A larger number of participants might have led to different results but also more robust findings. For example, interpreting the correlation results was hampered by the fact that, due to the small sample size, some statistically non-significant relationships nevertheless had correlation coefficients that, according to Plonsky and Oswald's (2014) benchmark guidelines, could be considered small or even medium effect sizes.

Also, although the student participants were all placed in the same low-intermediate level, it cannot be guaranteed that they all had the same level of speaking proficiency. Students were placed in this level based on their results on a standardized placement test (TOEFL ITP), which contained listening, reading, and writing elements but not a speaking component; therefore, it is likely that the classes were somewhat mixed in terms of their speaking ability. Furthermore, the lack of a control group means that no comparisons can be made with other proficiency levels or learning environments. It is therefore unclear to what extent the gains made in this study can be attributed to participation in the class activities. It is also unclear to what extent students enrolled in a different kind of course, or taught according to a different methodology, would have developed differently from those who participated in this course.

Another limitation concerns the quantification of interactional competence. Although the topic development and support moves were adapted and coded according to the specific context of the study, the contingencies of interaction mean that not all types of moves were required by each participant during each discussion. As a result, participants were not able to display the full range of their competence in every situation, as the appropriacy or relevance of specific moves was constrained by the behavior of other participants during each exchange. Therefore, conclusions about the development of individuals over time should be made cautiously. In addition, aggregating all topic development and support moves into two domains for the purpose of correlation analysis overlooks the inherent differences between categories. For example, although both were counted as instances of support, backchannels and repair moves are different types of features, relying on different cognitive processes, which again calls for interpretations to be made cautiously.

Finally, although there is a limit to the resources of an individual researcher, other measures and data-gathering techniques could have been used to supplement these findings. For instance, despite its methodological difficulties, attempting to measure the utterance fluency of individuals during the group discussions might have strengthened conclusions regarding whether fluency development displayed during the monologues was matched by development in the interactive context. Also, measuring the lexical and morpho-syntactic accuracy and complexity of learner output could have addressed the issue of whether any fluency gains were achieved at the expense of other aspects of speech production. On the qualitative side, in addition to the end of course questionnaire, interviews with some of the participants would have provided further insight into their attitude to the course, as well as which skills and activities they valued.

Suggestions for Future Research

The results of this study, and the limitations outlined above, suggest several possibilities for future research. First, the study could be replicated with a larger *N*-size. Including more participants would enable more robust conclusions to be drawn and might allow patterns of fluency and interactional competence development to be observed more clearly. Also, the MFRM data could be reanalyzed with the number of facets expanded to include time (i.e., Time 1, 2, and 3), and the two categories of fluency and communicative adequacy. This analysis would allow the difficulty ranking of fluency and communicative adequacy to be compared on the same logit scale.

Next, although there are benefits to ascertaining whether development occurs in a single semester, conducting the study over a longer period of time—such as a full academic year—would allow conclusions to be drawn about the continued trajectory of development. For example, it would be interesting to see whether the current study's findings of longer comments and increased turn taking represented a ceiling that was reached relatively quickly or whether this development continued in some form. Other patterns might take longer to emerge, especially in an EFL environment in which learners do not have regular exposure to the target language outside of class.

Replicating this research with learners of different proficiency levels would also be beneficial. It is possible that both fluency and interactional competence develop differently depending on the linguistic knowledge that participants have at the outset of a course of study; therefore, comparing findings from different levels could provide valuable insight into which aspects of development overlap and which are distinct. It would also be interesting to hear the perspective of learners from different proficiency levels, such as the degree to which higher-level learners also value the direct teaching of interactional strategies, or whether they feel constrained by them having already attained sufficient competence and flexibility to navigate various interactional contexts.

Similarly, comparing results from different faculties or L1 backgrounds could yield further insights. The participants in this study were members of a liberal arts faculty in which studying English is prioritized, with six 100-minute classes a week required for freshman students. If development was also observed among students from different faculties, who are perhaps less receptive to English study, it would reinforce the benefits of this pedagogical approach. It is also possible that L1 and related cultural factors could have influenced the results. For example, patterns of turn taking are unlikely to be identical among speakers of all languages, despite the underlying similarities pointed out previously. It would therefore be desirable to compare the results of learners from different L1 backgrounds.

Rather than replicating the study, another approach could be to vary task conditions and compare the results. For example, having learners discuss topics without being taught interactional strategies could function as a control group, thus highlighting the way that such skills develop without direct instruction, which would allow educators to focus more on the skills that appeared to require instruction. Alternatively, other conditions—such as story-telling or decision-making tasks—could be used in addition to opinion-based group discussions to evaluate a wider range of speaking abilities. Comparing results with those obtained from natural interactions in non-controlled environments, such as informal or business settings, would also be valuable.

Finally, conducting more fluency studies in a group context would also be beneficial, given the absence of research in this area. A combination of individual variables (i.e., fluent syllables produced per minute) and group variables (i.e., number and length of inter-turn gaps) were included in this study, none of which have been examined before to the best of my knowledge. Whether researchers choose to use these variables or adopt their own, more needs to be done to investigate fluency development in interactive contexts, and to compare the findings with those from more established areas of research, such as monologue-based fluency development.

Final Comments

I have spent many years teaching discussion-based speaking classes to Japanese university students, starting long before I became involved in academic research, and was struck from the very beginning by how positively most students seemed to feel about these classes. The students seemed to feel that they were both worthwhile and enjoyable, and I had a general impression that they were improving their speaking skills. After a while, however, I began to wonder whether this apparent progress could be ascertained more concretely and perhaps measured in some way. Thus began a multi-year journey that ultimately led to the current study. I now feel much closer to answering those questions and have a far better understanding of precisely how fluency and the ability to interact develops in this kind of communicative EFL environment, as well as how these elements of speaking proficiency relate to learners' feelings of self-efficacy.

From the learner perspective, I have often heard students complain about the lack of speaking opportunities afforded in formal education prior to university, as the emphasis in such settings is on passing entrance exams, which results in heavily form-focused lessons. Therefore, a meaning-focused discussion and speaking course—as part of a balanced curriculum, and with appropriate feedback provided—is one way to redress this imbalance. Furthermore, I believe that having the opportunity to express their opinions and put previously learned grammar and vocabulary into productive use, is a major reason why students value such lessons and develop a positive attitude towards them. A final comment from the end of course questionnaire sums up this perspective, "At first I was nervous, but I became more and more able to convey what I wanted to say and by the end it was really enjoyable."

For language practitioners, the benefits of student-centered, communicative environments are often held to be self-evident, even if actual evidence—beyond the kind of general impression I referred to above—is lacking. I hope this study goes some way to addressing this issue and providing empirical support for what many educators—and learners—intuitively believe. Given that many learners prioritize their speaking skills, it is important that language teachers and researchers find out more about the processes involved in cultivating these skills. I plan to continue with this line of research, and hope that others investigate similar points of interest in order to deepen the understanding of L2 speaking development.

REFERENCES

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

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*, 561–573. https://doi.org/10.1007/BF02293814

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

Artunc, E. K., & Hart, D. O. (2019). Interactional competence in paired speaking tests: A study in proficiency-based pairings. *System*, *89*, 1–14. https://doi.org/10.1016/j.system.2019.102194

Atkinson, J. M., & Heritage, J. (1984). Transcription notation. In J. M. Atkinson & J. Heritage (Eds.), *Structures of social action: Studies in conversation analysis*. (pp. ix–xvi). Cambridge University Press.

Bachman, L. F. (2007). What is the construct? The dialectic of abilities and contexts in defining constructs in language assessment. In J. Fox, M. Wesche, D. Bayliss, L. Cheng, C. E. Turner, & C. Doe (Eds.), *Language testing reconsidered* (pp. 41–71). University of Ottawa Press.

Baker-Smemoe, W., Dewey, D. P., Bown, J., & Martinsen, R. A. (2014). Does measuring L2 utterance fluency equal measuring overall L2 proficiency? Evidence from five languages. *Foreign Language Annals*, *47*(4), 707–728. https://doi.org/10.1011/flan.12110

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Prentice Hall.

Bandura, A. (1997). *Self-efficacy: The exercise of control*. W. H. Freeman.

Bandura, A. (2006). Guide to the construction of self-efficacy scales. In F. Pajares & T. Urdan (Eds.), *Self-efficacy beliefs of adolescents* (pp. 307–337). Information Age.

Baran-Łukarz, M. (2022). Language anxiety. In T. M. Derwing, M. J. Munro, & R. I. Thomson (Eds.), *The Routledge handbook of second language acquisition and speaking* (pp. 83–96). Routledge.

Bardovi-Harlig, K., & Salsbury, T. (2004). The organization of turns in the disagreements of L2 learners: A longitudinal perspective. In D. Boxer, & A. D. Cohen (Eds.), *Studying speaking to inform second language learning* (pp. 199–227). Multilingual Matters.

Bielak, J. (2022). To what extent are foreign language anxiety and foreign language enjoyment related to L2 fluency? An investigation of task-specific emotions and breakdown and speed fluency in an oral task. *Language Teaching Research*, *29*(3). https://doi.org/10.1177/13621688221079319

Boersma, P., & Weenink, D. (2021). Praat: Doing phonetics by computer [Computer program]. Version 6.2.01. http://www.praat.org

Bond, T. G., Yan, Z., & Heene, M. (2021). *Applying the Rasch model: Fundamental measurement in the human sciences* (4th ed.). Routledge.

Bosker, H. R., Pinget, A-F., Quene, H., Sanders, T., & De Jong, N. H. (2013). What makes speech sound fluent? The contributions of pauses, speed and repairs. *Language Testing*, *30*(2), 159–175. https://doi.org/10.1177/0265532212455394

Botes, E., Dewaele, J-M., & Greiff, S. (2020). The foreign language classroom anxiety scale and academic achievement: An overview of the prevailing literature and a meta-analysis. *Journal for the Psychology of Language Learning*, *2*, 26–56. https://doi.org/10.52598/jpll/2/1/3

Brouwer, C. E., & Wagner, J. (2004). Developmental issues in second language conversation. *Journal of Applied Linguistics and Professional Practice*, *1*(1), 29–47. https://doi.org/10.1558/japl.v1.i1.29

Brown, D., Fernandez, J., & Huensch, A. (2023). The role of pragmatic markers in perceptions of L2 fluency in dialogue. *System*, *119*, 1–13. https://doi.org/10.1016/j.system.2023.103157

Busse, V. & Walter, C. (2013). Foreign language motivation in higher education: A longitudinal study of motivational changes and their causes. *The Modern Language Journal*, *97*(2), 435–456. https://doi:10.1111/j.1540-4781.2013. 12004.x.

Bygate, M. (2001). Effects of task repetition in the structure and control of oral language. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks, second language learning, teaching, and testing* (pp. 23–48). Routledge.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, *1*, 1–47. https://doi.org/10.1093/applin/l.1.1

Cao, Y., & Philp, J. (2006). Interactional context and willingness to communicate: A comparison of behavior in whole class, group and dyadic interaction. *System*, *34*(4), 480–493. https://doi.org/10.1016/j.system.2006.05.002

Carroll, D. (2004). Restarts in novice turn beginnings: Disfluencies or interactional achievements? In R. Gardner & J. Wagner (Eds.), *Second language conversations* (pp. 201–220). Continuum.

Chalhoub-Deville, M. (2003). Second language interaction: Current perspectives and future trends. *Language Testing*, *20*(4), 369–383. https://doi.org/10.1191/0265532203lt264oa

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum.

Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). Sage.

Creswell, J. W., & Plano Clark, V. L. (2007). *Designing and conducting mixed methods research* (1st ed.). Sage.

Creswell, J. W., & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research* (3rd ed.). Sage.

De Bot, K. (1992). A bilingual production model: Levelt's 'speaking' model adapted. *Applied Linguistics*, *13*(1), 1–24. https://doi.org/10.1093/applin/13/1/1

De Bot, K., & Larsen-Freeman, D. (2011) Researching second language development from a dynamic systems theory perspective. In M. H. Verspoor, K. de Bot, & W. Lowie (Eds.), *A Dynamic Approach to Second Language Development* (pp. 5–23). Benjamins.

De Jong, N. H. (2016). Predicting pauses in L1 and L2 speech: The effects of utterance boundaries and word frequency. *International Review of Applied Linguistics in Language Teaching*, *54*(2), 39–57. https://doi.org/10.1515/iral-2016-9933

De Jong, N. H. (2018). Fluency in second language testing: Insights from different disciplines. *Language Assessment Quarterly*, *15*(3), 237–254. https://doi.org/10.1080/15434303.2018.1477780

De Jong, N. H., & Bosker, H. R. (2013). Choosing a threshold for silent pauses to measure second language fluency. In R. Eklund (Ed.), *Proceedings of the 6th Workshop on Disfluency in Spontaneous Speech (DiSS)* (pp. 17–20). Royal Institute of Technology (KTH).

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

De Jong, N. H., Steinel, M. P., Florijn, A., Schoonen, R., & Hulstijn J. H. (2012). Facets of speaking proficiency. *Studies in Second Language Acquisition*, *34*(1), 5–34. https://doi.org/10.1017/S0272263111000489

De Jong, N. H., Steinel, M. P., Florijn, A., Schoonen, R., & Hulstijn J. H. (2013). Linguistic skills and speaking fluency in a second language. *Applied Psycholinguistics*, *34*, 893–916. https://doi.org/10.1017/S0142716412000069

DeKeyser, R. (2007). *Practice in a second language: Perspectives from applied linguistics and cognitive psychology.* Cambridge University Press.

DeKeyser, R. (2010). Monitoring processes in Spanish as a second language during a study abroad program. *Foreign Language Annals, 43*(1), 80–92. https://doi.org/10.1111/j.1944-9720.2010.01061.x

DeKeyser, R., & Botana, G. P. (2019). Current research on instructed second language learning: A bird's eye view. In R. DeKeyser & G. P. Botana (Eds.), *Doing SLA research with implications for the classroom: Reconciling methodological demands and pedagogical applicability* (pp. 1–8). Benjamins.

Dewaele, J. -M. (2007). The effect of multilingualism and socio-situational factors on communicative anxiety and foreign language anxiety of mature language learners. *International Journal of Bilingualism*, *11*(4), 391–409. https://doi.org/10.1177/13670069070110040301

Dewaele, J. -M. (2013). The link between foreign language classroom anxiety and psychoticism, extraversion, and neuroticism among adult bi- and multilinguals. *The Modern Language Journal*, *97*(3), 670–684. https://doi.org/10.1111/j.1540-4781.2013.12036.x

Dewaele, J. -M., & Furnham, A. (1999). Extraversion: The unloved variable in applied linguistics research. *Language Learning*, *49*(3), 509–544. https://doi.org/10.1111/0023-8333.00098

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

Doe, T. (2025). Fluency development in an EFL setting: A one-semester study. *Language Teaching Research*, *29*(1), 363–384. https://doi.org/10.1177/13621688211058520

Dörnyei, Z (2010). *Questionnaires in second language research* (2nd ed.). Routledge.

Dörnyei, Z., & Murphey, T. (2003). *Group dynamics in the language classroom.* Cambridge University Press.

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

Dörnyei, Z., & Thurrell, S. (1994). Teaching conversational skills intensively: Course content and rationale. *ELT Journal*, *48*(1), 40–49. https://doi.org/10.1093/elt/48.1.40

Dörnyei, Z., & Ushioda, E. (2011). *Teaching and researching motivation* (2nd ed.). Pearson Education.

Ducasse, A. M., & Brown, A. (2009). Assessing paired orals: Raters' orientation to interaction. *Language Testing*, *26*(3), 423–443. https://doi.org/10.1177/0265532209104669

Eckes, T. (2015). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments* (2nd ed.). Lang.

Ellis, N. (2001). Memory for language. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 33–68). Cambridge University Press.

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

Elwood, J. (2011). *Enriching structural models of L2 willingness to communicate: The role of personality, ego permeability, and perceived distance* (Unpublished doctoral dissertation). Temple University Japan.

Eysenck, M. W. (1981). Learning, memory and personality. In H. J. Eysenck (Ed.), *A model for personality* (pp. 169–209). Springer Verlag.

Fillmore, C. J. (1979). On fluency. In C. J. Fillmore, D. Kempler, & W. S-Y. Wang (Eds.), *Individual differences in language ability and language behavior* (pp. 85–101). Academic Press.

Field, A. (2024). *Discovering statistics using SPSS* (6th ed.). Sage.

Field, J. (2011). Cognitive validity. In L. Taylor (Ed.), *Examining speaking: Research and practice in assessing second language speaking* (pp. 65–111). Cambridge University Press.

Firth, A., & Wagner, J. (1997). On discourse, communication, and (some) fundamental concepts in SLA research. *The Modern Language Journal*, *81*(3), 285–300. https://doi.org/10.2307/329302

Fisher, W. P. (2007). Rating scale instrument quality criteria. *Rasch Measurement Transactions*, *21*, 1095.

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

Freed, B. F. (2000). Is fluency, like beauty, in the eyes (and ears) of the beholder? In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 243–265). University of Michigan Press.

Freed, B. F., Segalowitz, N., & Dewey, D. P. (2004). Context of learning and second language fluency in French: Comparing regular classroom, study abroad, and intensive domestic immersion programs. *Studies in Second Language Acquisition*, *26*(2), 275–301. https://doi.org/1017/S0272263104262064

Fulcher, G. (2010). *Practical language testing*. Hodder Education.

Galaczi, E. D. (2008). Peer-peer interaction in a speaking test: The case of the First Certificate in English examination. *Language Assessment Quarterly*, *5*(2), 89–119. https://doi.org/10.1080/15434300801934702

Galaczi, E. D. (2014). Interactional competence across proficiency levels: How do learners manage interaction in paired speaking tests? *Applied Linguistics*, *35*(5), 553–574. https://doi.org/10.1093/applin/amt017

Galaczi, E. D., & Taylor, L. (2018). Interactional competence: Conceptualisations, operationalisations, and outstanding questions. *Language Assessment Quarterly*, *15*(3), 219–236. https://doi.org/10.1080/15434303.2018.1453816

Galaczi, E. D., & Taylor, L. (2020). Measuring interactional competence. In P. Winke, & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing* (pp. 338–348). Routledge.

Gan, Z. (2010). Interaction in group oral assessment: A case study of higher- and lower-scoring students. *Language Testing*, *27*(4), 585–602. https://doi.org/10.1177/0265532210364049

Gatbonton, E., & Segalowitz, N. (1988). Creative automatization: Principles for promoting fluency within a communicative framework. *TESOL Quarterly*, *22*(3), 473–492. https://doi.org/10.2307/3587290

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

Goetze, J., & Driver, M. (2022). Is learning really just believing? A meta-analysis of self-efficacy and achievement in SLA. *Studies in Second Language Learning and Teaching*, *12*(2), 233–259. https://doi.org/10.14746/sslt.2022.12.2.4

Grice, H. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and semantics* (pp. 41–58). Academic.

Hair, J. F., Hult, G. T. M., Ringle, C. M, & Sarstedt, M. (2022). *A primer on partial least squares structural equation modelling (PLS-SEM)* (3rd ed.). Sage.

Hall, J. K., (2018). From L2 interactional competence to L2 interactional repertoires: Reconceptualising the objects of L2 learning. *Classroom Discourse*, *9*(1), 25–39. https://doi.org/10.1080/19463014.2018.1400050

Hall, J. K., & Pekarek Doeler, S. (2011). L2 interactional competence and development. In J. K. Hall, J. Hellermann, & S. Pekarek Doehler (Eds.), *L2 interactional competence and development* (pp. 1–15). Multilingual Matters.

Hanzawa, K. (2021). Development of second language speech fluency in foreign language classrooms: A longitudinal study. *Language Teaching Research*, *28*(3), 1–23. https://doi.org/10.1177/13621688211008693

He, A., & Young, R. (1998). Language proficiency interviews: A discourse approach. In R. Young, & A. He (Eds.), *Talking and testing: Discourse approaches to the assessment of oral proficiency* (pp. 1–24). Benjamins.

Hellermann, J. (2007). The development of practices for action in classroom dyadic interaction: Focus on task openings. *The Modern Language Journal*, *91*(1), 83–96. https://doi.org/10.1111/j.1540-4781.2007.00503.x

Hellermann, J. (2008). *Social actions for classroom language learning.* Multilingual Matters.

Hellermann, J. (2009a). Looking for evidence of language learning in practices for repair: A case study of self-initiated self-repair by an adult learner of English. *Scandinavian Journal of Educational Research*, *53*(2), 113–132. https://doi.org/10.1080/00313830902757550

Hellermann, J. (2009b). Practices for dispreferred responses using *no* by a learner of English. *International Review of Applied Linguistics in Language Teaching*, *47*(1), 95–126. https://doi.org/10.1515/iral.2009.005

Hellermann, J. (2011). Members' methods, members' competencies: Looking for evidence of language learning in longitudinal investigations of other-initiated repair. In J. K. Hall, J. Hellermann, & S. Pekarek Doehler (Eds.), *L2 interactional competence and development* (pp. 147–173). Multilingual Matters.

Hewitt, E., & Stephenson, J. (2012). Foreign language anxiety and oral exam performance: A replication of Phillips's *MLJ* study. *The Modern Language Journal*, *96*(2), 170–189. https://doi.org/10.1111/j.1540-4781.2011.01174.x

Holt, E., & Drew, P. (2005). Figurative pivots: The use of figurative expressions in pivotal topic expressions. *Research on Language and Social Interaction*, *38*(1), 35–61. https://doi.org/10.1207/s15327973rlsi3801_2

Horwitz, E. K. (2010). Foreign and second language anxiety. *Language Teaching*, *43*(2), 154–167. https://doi.org/10.1017/S026144480000036X

Horwitz, E. K., Horwitz, M. B., & Cope, J. (1986). Foreign language classroom anxiety. *The Modern Language Journal*, *70*(2), 125–132. https://doi.org/10.2307/327317

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. https://doi.org/10.1093/applin/amm017

Iwashita, N., May, L., & Moore, P. J. (2021). Operationalizing interactional competence in computer-mediated speaking tests. In M. R. Salaberry & A. R. Burch (Eds.), *Assessing speaking in context* (pp. 283–302). Multilingual Matters.

Kahng, J. (2014). Exploring utterance and cognitive fluency of L1 and L2 English speakers: Temporal measures and stimulated recall. *Language Learning*, *64*(4), 809–854. https://doi.org/10.1111/lang.12084

Kahng, J. (2020). Explaining second language utterance fluency: Contribution of cognitive fluency and first language utterance fluency. *Applied Psycholinguistics*, *41*, 457–480. https://doi.org/10.1017/S0142716420000065

Kahng, J. (2024). Longitudinal development of second language utterance fluency, cognitive fluency, and their relationship. *Studies in Second Language Acquisition*, *46*, 535–549. https://doi.org/10.1017/S027226312300591

Kasper, G. (2006). Beyond repair: Conversation analysis as an approach to SLA. *AILA Review*, *19*, 83–99. https://doi.org/1075/AILA.19.07KAS

Kasper, G. (2009). Locating cognition in second language interaction and learning: Inside the skull or in public view? *International Review of Applied Linguistics*, *47*, 11–36. https://doi.org/10.1515/iral.2009.002

Kecskes, I., Sanders, R. E., & Pomerantz, A. (2018). The basic interactional competence of language learners. *Journal of Pragmatics*, *124*, 88–105. https://doi.org/10.1016/j.pragma.2017.10.019

Kley, K., Kunitz, S., & Meng, Y. (2021). Jiazhou? Is it California? Operationalizing repair in classroom-based assessment. In M. R. Salaberry & A. R. Burch (Eds.), *Assessing speaking in context* (pp. 165–191). Multilingual Matters.

Koizumi, R., Agawa, T., Asano, K., & In'nami, Y. (2022). Skill profiles of Japanese English learners and reasons for uneven patterns. *Language Testing in Asia*, *12*(1), 1–34. https://doi.org/10.1186/s40468-022-00203-3

Koponen, M., & Riggenbach, H. (2000). Overview: Varying perspectives on fluency. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 5–24). University of Michigan Press.

Kormos, J. (1999). Monitoring and self-repair in L2. *Language Learning*, *49*(2), 303–342. https://doi.org/10.1011/0023-8333.00090

Kormos, J. (2006). *Speech production and second language acquisition*. Routledge.

Kormos, J., & Denes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System*, *32*, 145–164. https://doi.org/10.1016/j.system.2004.01.001

Kramsch, C. (1986). From language proficiency to interactional competence. *The Modern Language Journal*, *70*(4), 366–372. https://doi.org/10.1111/modl.1986.70.issue-4

Lam, D. M. K. (2018). What counts as 'responding'? Contingency on previous speaker contribution as a feature of interactional competence. *Language Testing*, *35*(3), 377–411. https://doi.org/10.1177/0265532218758126

Lam, D. M. K., Galaczi, E., Nakatsuhara, F., & May, L. (2023). Assessing interactional competence: Exploring ratability challenges. *Applied Pragmatics, 5*(2), 208–233. https://doi.org/10.1075/ap.00014.lam

Lambert, C., Kormos, J, & Minn, D (2017). Task repetition and second language speech processing. *Studies in Second Language Acquisition*, *39*(1), 167–196. https://doi.org/10.1017/S0272263116000085

Leonard, R. K., & Shea, C. E. (2017). L2 speaking development during study abroad: Fluency, accuracy, complexity, and underlying cognitive factors. *The Modern language Journal*, *101*(1), 179–193. https://doi.org/10.1011/modl.12382

Lee, Y. -A. (2006). Towards respecification of communicative competence: Condition of L2 instruction or its objective? *Applied Linguistics*, *27*(3), 349–376. https://doi.org/10.1093/applin/aml011

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*, 1–18. https://doi.org/10.1186/s40862-017-0035-x

Leeming, P., Vitta, J. P., Hiver, P., Hicks, D., McLean, S., & Nicklin, C. (2024). Willingness to communicate, speaking self-efficacy, and perceived communicative competence as predictors of second language spoken task production. *Language Learning*, *74*(4), 917–949. https://doi.org/10.1111/lang.12640

Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning*, *40*(3), 387–417. https://doi.org/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). University of Michigan Press.

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

Levelt, W. J. M. (1999). Language production: A blueprint of the speaker. In C. Brown & P. Hagoort (Eds.), *Neurocognition of language* (pp. 83–122). Oxford University Press.

Levinson, S. C. (2006). On the human "interaction engine." In N. J. Enfield & S. C. Levinson (Eds.), *Roots of human sociality: Culture, cognition and interaction* (pp. 39–69). Berg.

Levinson, S. C. (2022). The interaction engine: Cuteness selection and the evolution of the interactional base for language. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *377*, 1–7. https://doi.org/10.1098/rstb.2021.0108

Lieberman, M. D. (2000). Introversion and working memory: Central executive differences. *Personality and Individual Differences*, *28*, 479–486. https://doi.org/10.1016/S0191-8869(99)00113-0

Lightbown, P. M. (2008). Transfer appropriate processing as a model for classroom second language acquisition. In Z. Han (Ed.), *Understanding second language process* (pp. 27–44). Multilingual Matters.

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

Linacre, J. M. (2002b). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, *16*, 878.

Linacre, J. M. (2017). A user's guide to WINSTEPS: Rasch-model computer programs (4.0.0) [Computer software manual]. Winsteps.com.

Linacre, J. M. (2023). WINSTEPS Rasch measurement computer program (Version 5.6.0). [Computer software]. Winsteps.com

Linacre, J. M. (2024). FACETS Rasch measurement computer program (Version 4.1.4). [Computer software]. Winsteps.com

Logan, G. D. (1988). Toward in instance theory of automatization. *Psychological Review*, *95*(4), 492–527. https://doi.org/10.1037/0033-295X.95.4.492

Long, M. (1998). Focus on form in task-based language teaching. *University of Hawai'i Working Papers in ESL*, *16*, 35–49. https://doi.org/10.1075/z.96.11lon

MacIntyre, P. D., & Gardner, R. C. (1989). Anxiety and second-language learning: Toward a theoretical clarification. *Language Learning*, *59*(4), 589–608. https://doi.org/10.1111/j.1467-1770.1989.tb00423.x

MacIntyre, P. D., & Wang, L. (2022). Anxiety. In S. Li, P. Hiver, & M. Papi (Eds.), *The Routledge handbook of second language acquisition and individual differences* (pp.175–189). Routledge.

McAdams, D. P., & Pals, J. L. (2006). A new big five: Fundamental principles for an integrative science of personality. *American Psychologist*, *61*(3), 204–217. https://doi.org/10.1037/0003-066X.61.3.204

McCarthy, M. (2010). Spoken fluency revisited. *English Profile Journal*, *1*(1), 1–15. https://doi.org/10.1017/S2041536210000012

Michel, M. C. (2011). Effects of task complexity and interaction on L2 performance. In P. Robinson (Ed.), *Second language task complexity: Researching the cognition hypothesis of language learning and performance* (pp. 141–174). Benjamins.

Mills, N. A. (2014). Self-efficacy in second language acquisition. In S. Mercer & M. Williams (Eds.), *Multiple perspectives on the self in SLA* (pp. 239–266). Multilingual Matters.

Mills, N. A., Pajares, F., & Herron, C. (2007). Self-efficacy of college intermediate French students: Relation to achievement and motivation. *Language Learning*, *57*(3), 417–442. https://doi.org/10.1111/j.1467-9922.2007.00421.x

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

Muñoz, C. (2008). Symmetries and asymmetries of age effects in naturalistic and instructed L2 learning. *Applied Linguistics*, *29*(4), 578–596. https://doi.org/10.1093/applin/amm056

Nakatani, Y. (2010). Identifying strategies that facilitate EFL learners' oral communication: A classroom study using multiple data collection procedures. *The Modern Language Journal*, *94*(1), 116–136. https://doi.org/10.1111/j.1540-4781.2009.00987.x

Nakatsuhara, F. (2011). Effects of test-taker characteristics and the number of participants in group oral tests. *Language Testing*, *28*(4), 483–508. https://doi.org/10.1177/0265532211398110

Nakatsuhara, F. (2013). *The co-construction of conversation in group oral tests*. Peter Lang.

Nation, I. S. P. (1989). Improving speaking fluency. *System*, *17*, 377–384. https://dx.doi.org/10.1016/0346-251X(89)90010-9

Nation, I. S. P. (2007). The four strands. *Innovation in Language Learning and Teaching*, *1*(1), 2–13. https://doi.org/10.2167/illt.039.0

Nation, I. S. P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge University Press.

Nishino, T., & Watanabe, M. (2008). Communication-oriented policies versus classroom realities in Japan. *TESOL Quarterly*, *42*(1), 133–138. https://doi.org/10.2307/40264432

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. https://doi.org/10.1177/0265532213514401

Ockey, G. J. (2011). Self-consciousness and assertiveness as explanatory variables of L2 Oral Ability: A latent variable approach. *Language Learning*, *61*(3), 968–989. https://doi.org/10.1111/j.1467-9922.2010.00625.x

Ockey, G. J., Koyama, D., Setoguchi, E., & Sun, A. (2015). The extent to which TOEFL iBT speaking scores are associated with performance on oral language tasks and oral ability components for Japanese university students. *Language Testing*, *32*(1), 39–62. https://doi.org/10.1177/0265532214538014

O'Connor, M. C., & Paunonen, S. V. (2007). Big five personality predictors of post-secondary academic performance. *Personality and Individual Differences*, *43*, 971–990. https://doi.org/10.1016/j.paid.2007.03.017

Pajares, F. (1996). Self-efficacy beliefs in academic settings. *Review of Educational Research*, *66*, 543–578. https://doi.org/10.3102/00346543066004543

Pajares, F., & Schunk, D. H. (2001). Self-beliefs and school success: Self-efficacy, self-concept, and school achievement. In R. Riding & S. Raynor (Eds.), *Perception* (pp. 239–266). Ablex.

Pallotti, G. (2009). CAF: Defining, refining and differentiating constructs. *Applied Linguistics*, *30*(4), 590–601. https://doi.org/10.1093/applin/amp045

Pallotti, G. (2019). Assessing tasks: The case of interactional difficulty. *Applied Linguistics*, *40*(1), 176–197. https://doi.org/10.1093/applin/amx020

Pallotti, G. (2020). Measuring complexity, accuracy, and fluency (CAF). In P. Winke & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing* (pp. 201–210). Routledge.

Pan, C., & Zhang, X. (2023). A longitudinal study of foreign language anxiety and enjoyment. *Language Teaching Research*, *27*(6), 1552–1575. https://doi.org/10.1177/1362168821993341

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

Pekarek Doehler, S. (2018). Elaborations on L2 interactional competence: The development of L2 grammar-for-interaction. *Classroom Discourse*, *9*(1), 3–24. https://doi.org/10.1080/19463014.2018.1437759

Pekarek Doehler, S. (2019). On the nature and the development of L2 interactional competence: State of the art and implications for praxis. In M. R. Salaberry & S. Kunitz (Eds.), *Teaching and testing L2 interactional competence* (pp. 25–59). Routledge.

Pekarek Doehler, S., & Pochen-Berger, E. (2011). Developing 'methods' for interaction: A cross-sectional study of disagreement sequences in L2 French. In J. K. Hall, J. Hellermann, & S. Pekarek Doehler (Eds.), *L2 interactional competence and development* (pp. 206–243). Multilingual Matters.

Pekarek Doehler, S., & Pochen-Berger, E. (2015). The development of L2 interactional competence: Evidence from turn-taking organization, sequence organization, repair organization, and preference organization. In T. Cadierno & S. W. Eskildsen (Eds.), *Usage-based perspectives in second language learning* (pp. 233–268). De Gruyter Mouton.

Peltonen, P. (2017). Temporal fluency and problem-solving in interaction: An exploratory study of fluency resources in L2 dialogue. *System*, *70*, 1–13. https://doi.org/10.1016/j.system.2017.08.009

Pérez Castillejo, S. (2019). The role of foreign language anxiety on L2 utterance fluency during a final exam. *Language Testing*, *36*(3), 327–345. https://doi.org/10.1177/0265532218777783

Phipps, J. (2024). A mixed-method analysis of self-efficacy for speaking with university EFL learners. *The Language Learning Journal*, *52*(6), 677–693. https://doi.org/10.1080/09571736.2023.2227866

Phillips, E. M. (1992). The effects of language anxiety on students' oral test performance and attitudes. *The Modern Language Journal*, *76*(1)*,* 14–26. https://doi.org/10.2307/329894

Plonsky, L. & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*(4), 878–912. https://doi.org/10.1111/lang.12079

Pomerantz, A. (1984). Agreeing and disagreeing with assessments: Some features of preferred/dispreferred turn shapes. In J. Atkinson & J. Heritage (Eds.), *Structures of social action* (pp. 57–101). Cambridge University Press.

Prefontaine, Y. (2013). Perceptions of French fluency in second language speech production. *The Canadian Modern Language Review*, *69*(3), 324–348. https://doi.org/10.3138/cmlr.1748

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

Révész, A., Ekiert, M., & Torgersen, E. (2016). The effects of complexity, accuracy, and fluency on communicative adequacy in oral task performance. *Applied Linguistics*, *37*(6), 828–848. https://doi.org/10.1093/applin/amu069

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

Riggenbach, H. (1991). Toward an understanding of fluency: A microanalysis of nonnative speaker conversations. *Discourse Processes*, *14*, 423–441. https://doi.org/10.1080/01638539109544795

Robinson, P. (2005). Cognitive complexity and task sequencing: Studies in a componential framework for second language task design. *International Review of Applied Linguistics*, *43*, 1–32. https://doi.org/10.1515/iral.2005.43.1.1

Roever, C. & Ikeda, N. (2021). What scores from monologic speaking tests can(not) tell us about interactional competence. *Language Testing*, *39*(1), 7–29. https://doi.org/10.1177/02655322211003332

Roever, C., & Kasper, G. (2018). Speaking in turns and sequences: Interactional competence as a target construct in testing speaking. *Language Testing*, *35*(3), 331–355. https://doi.org/10.1177/0265532218758128

Ross, S. (2018). Listener response as a facet of interactional competence. *Language Testing*, *35*(3), 357–375. https://doi.org/10.1177/0265532218758125

Rossiter, M. J., Derwing, T. M., Manimtim, L. G., & Thomson, R. I. (2010). Oral fluency: The neglected component in the communicative language classroom. *The Canadian Modern Language Review*, *66*(4), 583–606. https://doi.org/10.3138/cmlr.66.4.583

Ryan, S. (2009). Ambivalence and commitment, liberation and challenge: Investigating the attitudes of young Japanese people towards the learning of English. *Journal of Multilingual and Multicultural Development*, *30*, 405–420. https://doi.org/10.1080/01434630902928447

Salaberry, M. R., & Burch A. R. (2021). Assessing speaking in context: Expanding the construct and the applications. In M. R. Salaberry & A. R. Burch (Eds.), *Assessing speaking in context* (pp. 2–22). Multilingual Matters.

Salaberry, M. R., & Kunitz S. (2019). Introduction. In M. R. Salaberry & S. Kunitz (Eds.), *Teaching and testing L2 interactional competence* (pp. 1–22). Routledge.

Schegloff, E. A. (1993). Reflections on quantification in the study of conversation. *Research on Language and Social Interaction*, *26*(1), 99–128. https://doi.org/10.1207/s15327973rlsi2601_5

Schunk, D. H. (1991). Self-efficacy and academic motivation. *Educational Psychologist*, *26*, 207–231. https://doi.org/10.1080/00461520.1991.9653133

Seedhouse, P. (2004). *The interactional architecture of the language classroom: A conversation analysis perspective.* Blackwell.

Seedhouse, P. (2005). Conversation analysis and language learning. *Language Teaching*, *38*, 165–187. https://doi.org/10.1017/S0261444805003010

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

Segalowitz, N. (2010). *Cognitive bases of second language fluency.* Routledge.

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

Segalowitz, N., & Segalowitz, S. J. (1993). Skilled performance, practice, and the differentiation of speed-up from automatization effects: Evidence form second language word recognition. *Applied Psycholinguistics*, *14*, 369–385. https://doi.org/10.1017/S0142716400010845

Sert, O. (2019). The interplay between collaborative turn sequences and active listenership: Implications for the development of L2 interactional competence. In M. R. Salaberry & S. Kunitz (Eds.), *Teaching and testing L2 interactional competence* (pp. 142–166). Routledge.

Siegel, A. & Seedhouse, P. (2024). *Human spoken interaction as a complex adaptive system.* Edinburgh University Press.

Simsek, E., & Dörnyei, Z. (2017). Anxiety and L2 self-images: The 'anxious self'. In C. Gkonou, M. Daubney, & J.-M. Dewaele (Eds.), *New insights into language anxiety: Theory, research and educational implications* (pp. 51–69). Multilingual Matters.

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

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics*, *30*(4), 510–532. https://doi.org/10.1093/applin/amp047

Skehan, P., & Foster, P. (1999). The influence of task structure and processing conditions on narrative retellings. *Language Learning*, *49*(1), 93–120. https://doi.org/10.1111/1467-9922.00071

Skehan, P., Foster, P., & Shum, S. (2016). Ladders and snakes in second language fluency. *International Review of Applied Linguistics in Language Teaching*, *54*(2), 97–111. https://doi.org/10.1515/iral-2016-9992

Spielberger, C. D. (1983). *Manual for the state-trait anxiety inventory: STAI (form Y).* Consulting Psychologists Press.

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., … Levinson, S. C. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, *106*, 10587–10592. https://doi.org/10.1073/pnas.0903616106

Storch, N. (2002). Patterns of interaction in ESL pair work. *Language Learning*, *52*(1), 119–158. https://doi.org/10.1111/lang.2002.52.issue-1

Suzuki, S., & Kormos, J. (2020). Linguistic dimensions of comprehensibility and perceived fluency: An investigation of complexity, accuracy, and fluency in second language argumentative speech. *Studies in Second Language Acquisition*, *42*(1), 143–167. https://doi.org/10.1017/S0272263119000421

Suzuki, S., & Kormos, J. (2023). The multidimensionality of second language oral fluency: Interfacing cognitive fluency and utterance fluency. *Studies in Second Language Acquisition*, *45*(1), 38–64. https://doi.org/10.1017/S0272263121000899

Suzuki, S., Kormos, J., & Uchihara, T. (2021). The relationship between utterance and perceived fluency: A meta-analysis of correlational studies. *The Modern Language Journal*, *105*(2), 435–463. https://doi.org/10.1111/modl.12706

Taguchi, N. (2014). Development of interactional competence in Japanese as a second language: Use of incomplete sentences as interactional resources. *The Modern Language Journal*, *98*(2), 518–535. https://doi.org/10.1111/j.1540-4781.2014.12087.x

Tavakoli, P. (2011). Pausing patterns: Differences between L2 speakers and native speakers. *ELT Journal*, *65*(1), 71–79. https://doi.org/10.1093/elt/ccq020

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. https://doi.org/10.1515/iral-2016-9994

Tavakoli, P. (2023). Making fluency research accessible to second language teachers: The impact of a training intervention. *Language Teaching Research*, *27*(2), 368–393. https://doi.org/10.1177/1362168820951213

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

Tavakoli, P., Campbell, C., & McCormack, J. (2016). Development of speech fluency over a short period of time: Effects of pedagogic intervention. *TESOL Quarterly*, *50*(2), 447–471. https://doi.org/10.1002/tesq.244

Tavakoli, P. & Hunter, A.-M. (2018). Is fluency being 'neglected' in the classroom? Teacher understanding of fluency and related classroom practices. *Language Teaching Research*, *22*(3), 330–349. https://doi.org/10.1177/1362168817708462

Tavakoli, P., & Uchihara, T. (2019). To what extent are multiword sequences associated with oral fluency? *Language Learning*, *70*(2), 506–547. https://doi.org/10.1111/lang.12384

Tavakoli, P., Nakatsuhara, F., & Hunter, A.-M. (2020). Aspects of fluency across assessed levels of speaking proficiency. *The Modern Language Journal, 104*(1), 169–191. https://doi.org/10.1111/modl.12620

Tavakoli, P., & Wright, C. (2020). *Second language speech fluency: From research to practice.* Cambridge University Press.

Teimouri, Y., Goetze, J., & Plonsky, L. (2019). Second language anxiety and achievement: A meta-analysis. *Studies in Second Language Acquisition*, *41*(2), 275–301. https://doi.org/10.1017/S0272263118000311

Thai, C., & Boers, F. (2016). Repeating a monologue under increasing time pressure: Effects on fluency, complexity, and accuracy. *TESOL Quarterly*, *50*(2), 369–393. https://doi.org/10.1002/tesq.232

Towell, R., Hawkins, & R., Bazergui, N. (1996). The development of fluency in advanced learners of French. *Applied Linguistics*, *17*(1), 84–119. https://doi:10.1093 /applin/17.1.84

Turner, J. L. (2014). *Using statistics in small-scale language education research.* Routledge.

Van Batenburg, E. S. L., Oostdam, R. J., Van Gelderen, A. J. S., & De Jong, N. H. (2018). Measuring L2 speakers' interactional ability using interactive speech tasks. *Language Testing*, *35*(1)*,* 75–100. https://doi.org/10.1177/0265532216679452

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), 519–535. https://doi.org/10.1111/modl.12545

Van Os, M., De Jong. N. H., & Bosker, H. R. (2020). Fluency in dialogue: The effect of turn-taking behavior on perceived fluency in native and non-native speech, *Language Learning*, *70*(4), 1183–1217. https://doi.org/10.1111/lang.12416

Wagner, J., Pekarek Doehler, S., & González-Martínez, E. (2018). Longitudinal research on the organization of social interaction: Current developments and methodological challenges. In S. Pekarek Doehler, J. Wagner, & E. González-Martínez (Eds.), *Longitudinal research on the organization of social interaction* (pp. 3–35). Palgrave Macmillan.

Waring, H. Z. (2018). Teaching L2 interactional competence: Problems and possibilities. *Classroom Discourse*, *9*(1), 57–67. https://doi.org/10.1080/19463014.2018.1434082

Waring, H. Z. (2019). Developing interactional competence with limited linguistic resources. In M. R. Salaberry, & S. Kunitz (Eds.), *Teaching and testing L2 interactional competence: Bridging theory and practice* (pp. 215–227). Routledge.

Watanabe, Y., & Swain, M. (2007). Effects of proficiency differences and patterns of pair interaction second language learning: Collaborative dialogue between adult ESL learners. *Language Teaching Research*, *11*, 121–142. https://doi.org/10.1177/136216880607074599

Wei, L., Lin, H.-H., & Litton, F. (2018). Communicative Language Teaching (CLT) in EFL Context in Asia. *Asian Culture and History, 10*(2), 1–9. https://doi.org/10.5539/ach.v10n2p1

Wolfe, E. W., & McVay, A. (2012). Application of latent trait models to identifying substantively interesting raters. *Educational Measurement: Issues and Practice*, *31*(3), 31–37. https://doi.org/10.1111/j.1745-3992.2012.00241.x

Wolfe, E. W., & Smith, E. V. (2007). Instrument development tools and activities for measure validation using Rasch models: Part II–Validation activities. In E. V. Smith & R. M. Smith (Eds.), *Rasch measurement: Advanced and specialized applications* (pp. 243–290). JAM Press.

Wong, J, & Waring, H. Z. (2021). *Conversation analysis and second language pedagogy: A guide for ESL/EFL teachers* (2nd ed.). Routledge.

Wood, D. (2009). Effects of focused instruction of formulaic sequences on fluent expression in second language narratives: A case study. *Canadian Journal of Applied Linguistics*, *12*, 39–56. https://journals.lib.unb.ca/index.php/CJAL/article/view/19898

Wood, D. (2010). *Formulaic language and second language speech fluency*. Continuum.

Woodrow, L. (2006). Anxiety and speaking English as a second language. *Regional Language Centre Journal*, *37*(3), 308–328. https://doi.org/10.1177/0033688206071315

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge University Press.

Wray, A., & Perkins, M. R. (2000). The functions of formulaic language: An integrated model. *Language and Communication*, *20*(1), 1–28. https://doi.org/10.1016/S0271-5309(99)00015-4

Yan, X. (2020). Unpacking the relationship between formulaic sequences and speech fluency on elicited imitation tasks: Proficiency level, sentence length, and fluency dimensions. *TESOL Quarterly*, *54*(2), 460–487.

Young, R. F. (2008). *Language and interaction: An advanced resource book*. Routledge.

Young, R. F. (2011). Interactional competence in language learning, teaching, and testing. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (Vol. 2, pp. 426–443). Routledge.

Young, R. F., & Miller, E. R. (2004). Learning as changing participation: Discourse roles in ESL writing conferences. *The Modern Language Journal*, *88*(4), 519–535. https://doi.org/10.1111/j.0026-7902.2004.t01-16-.x

Zareva, A., Schwanenflugel, P, & Nikolova, Y. (2005). Relationship between lexical competence and language proficiency. *Studies in Second Language Acquisition*, *27*(4), 567–595. https://doi.org/10.1017/S0272263105050254

Zhang, X. (2019). Foreign language anxiety and foreign language performance: A meta-analysis. *The Modern Language Journal*, *103*(4), 763–781. https://doi.org/10.1111/modl.12590

Zhou, S., Chui, M. M, Dong, Z., & Zhou W. (2023). Foreign language anxiety and foreign language self-efficacy: A meta-analysis. *Current Psychology*, *42*, 31536–31550. https://doi.org/10.1007/s12144-022-04110-x

APPENDICES

APPENDIX A

INFORMED CONSENT (ENGLISH VERSION)

You are being invited to take part in a research study. A person who takes part in a research study is called a research subject, or research participant.

What should I know about this research?

- Someone will explain this research to you.
- This form sums up that explanation.
- Taking part in this research is voluntary. Whether you take part is up to you.
- You can choose not to take part. There will be no penalty or loss of benefits to which you are otherwise entitled.
- You can agree to take part and later change your mind. There will be no penalty or loss of benefits to which you are otherwise entitled.
- If you don't understand, ask questions.
- Ask all the questions you want before you decide.

Why is this research being done?

The purpose of this research is to better understand the development of Japanese university students' English speaking. Specifically, we are researching how students' fluency and interactional skills change over time. We are also looking at how students' feelings of self-efficacy and anxiety change during a speaking course. Understanding these changes is important to improve English teaching and testing techniques. This research will help English teachers to improve their speaking instruction and also help testers to score speaking tests more accurately.
About 40 subjects will take part in this research.

How long will I be in this research?

We expect that your taking part in this research will last for 3 months, for one academic semester.

What happens to me if I agree to take part in this research?

Being recorded is optional and is not a graded part of your class.

Schedule

- You will complete a background questionnaire at the start of the research.
- You will complete three 2-minute monologues, which will be audio recorded.
- You will complete three 9-minute (or 12-minute) group discussions, which will be video recorded.

- You will complete three further questionnaires regarding how you feel when speaking English in class.

About the questionnaires

The background questionnaire will take about 10 minutes to complete. You will be asked for some personal information such as your name, email, university name, department name, major, age, gender, and first language. You will also be asked some questions about your English ability. For example, about your English proficiency, vocabulary ability, and your recent scores on the TOEIC, TOEFL, etc. If you haven't taken a test or don't know the answer, you can leave the question blank.

The further questionnaires will take about 2 minutes to complete each time and will be conducted before each recording session. You will be asked to select answers relating to how you feel when speaking English in class.

About the speaking tasks

The monologue speaking tasks will take 2 minutes each to complete. The group discussions will take 9 minutes to complete if there are 3 members, and 12 minutes to complete if there are 4 members. Both the monologue and group discussion activities are similar to activities that you complete every week in class, including those which you are assessed on during speaking tests, so it is a good opportunity to practice for those tests. You will be given a topic and asked to talk about that topic.

> *Example monologue topic:*
> Say as much as you can about your favorite season (activities, weather, fashion, food, trips, etc.)
> *Example discussion questions:*
> 1. What is the best season in Japan?
> 2. Which do you think is better, summer or winter?

You will have to complete the tasks using English only and will not be able to ask the instructor for help during the recordings.
The researchers will study your speaking to learn more about how students' English speaking changes over time. At the end of the study, we will email you a report about the general findings if you request one.

What are my responsibilities if I take part in this research?

If you take part in this research, you will be responsible for completing the background questionnaire at the beginning of the research as well as three monologues, three group discussions, and three further questionnaires during the semester.

Could being in this research hurt me?

- There is a possible risk of discomfort due to completing the speaking tasks if you do not enjoy English speaking activities or being recorded while speaking.
- There is a possible privacy risk of you being identified by the other researchers who will listen to your speaking. The other researchers will not be given your real names and will have to undergo training in research ethics before watching your videos.

Will it cost me money to take part in this research?

Taking part in this research will not result in additional costs to you.

Will being in this research benefit me?

We cannot promise any benefits to you from your taking part in this research. However, possible benefits to you include improving your English speaking ability and having the opportunity to practice for future class and test speaking activities. The tasks used in this research are also a good way to practice and improve your English speaking fluency, so your English ability might improve if you participate in this research. Possible benefits to others include helping researchers to better understand how English speaking develops for students learning English as a foreign language. A better understand might benefit researchers, teachers, English program organizers, and test makers.

What happens to the information collected for this research?

Your private information may be shared with individuals and organizations (if applicable) that conduct or watch over this research, including:

- The researchers
- People who work with the researchers
- The Institutional Review Board (IRB) that reviewed this research
- Temple University

We may publish the results of this research. However, we will keep your name and other identifying information confidential. We protect your information from disclosure to others to the extent required by law. We cannot promise complete secrecy.

Who can answer my questions about this research?

If you have questions, concerns, or complaints, or think this research has hurt you or made you sick, talk to the research team at the phone number listed above on the first page.
This research is being overseen by an Institutional Review Board ("IRB"). An IRB is a group of people who perform independent review of research studies. You may talk to them at (215) 707-3390 or irb@temple.edu if:

- You have questions, concerns, or complaints that are not being answered by the research team.
- You are not getting answers from the research team.
- You cannot reach the research team.
- You want to talk to someone else about the research.
- You have questions about your rights as a research subject.

Can I be removed from this research without my approval?

The person in charge of this research can remove you from this research without your approval. Possible reasons for removal include:

- You do not complete the background questionnaire.
- It appears that you did not complete the speaking tasks in earnest, for example completely off-topic responses.
- You did not complete all of the speaking tasks.

We will tell you about any new information that may affect your health, welfare, or choice to stay in this research.

What happens if I agree to be in this research, but I change my mind later?

If you decide to leave this research, contact the research team so that the investigator can remove all or part of your data from the research upon your request. After leaving this research, you will still be allowed to complete the speaking tasks as a part of normal classroom activities. There will not be any penalty for withdrawing from the research.

Your consent to participate in this study

I understand the information given to me and agree to the conditions of this study.

____
Name (kanji)

____________________________________ __________________
Name (Romaji) Date

APPENDIX B

INFORMED CONSENT (JAPANESE VERSION)

あなたは研究に参加するように招かれています。調査研究に参加する人のことを研究対象者、または研究参加者といいます。

この研究について知っておくべきことは？

- この研究についての説明があります。
- この研究への参加は任意です。参加するかどうかはあなた次第です。
- 参加しなくても不利になることはありません。
- 参加して後でやめることになっても、それであなたが不利になることはありません。
- わからないことがあるときは質問してください。
- 決める前に聞きたいことを全部聞いてください。

なぜこの研究が行われているのか？

本研究の目的は、日本の大学生の英語の話力の発達をより一層理解することです。具体的には学生の流暢さと対話力が時間の経過とともにどのように変化するかを研究しています。また、学生の自己効力感と不安がどのように変化するかを研究しています。これらの変化を理解することは、英語の教育や試験の技術を向上させるために重要です。この研究は、英語教師が指導を改善するのに役立ち、試験作成者がスピーキングをより正確に採点するのにも役立ちます。

約 40 人がこの研究に参加します。

どのくらいの期間、この研究に参加しているのでしょうか？

この研究に参加する場合、一つの学期（約3か月）続くことを想定しています。

この研究に参加することに同意したら、私はどうなるのでしょうか？

この研究の録画はオプションであり、採点されるものではありません。

日程

- 調査の開始時にバックグラウンドアンケートに回答します。
- ２分かかるモノローグを 3 回録音します。
- ９分（や１２分）かかるグループディスカッションを 3 回録画します。
- 英語で話している感覚を調査するため、別のアンケートを 3 回実施します。

アンケートについて

バックグラウンドアンケートは 10 分程度で終わります。氏名、電子メール、大学名、学科名、専攻、年齢、性別、第一言語（母語）などの個人情報が質問されます。また、あなたの英語力についていくつか質問されます。例えば、英語能力、語彙力、最近の TOEIC、TOEFL の得点などについて質問されます。テストを受けたことがない場合や答えがわからない場合は、質問の回答

は空白のままで構いません。別のアンケートは録画の前に実施して、毎回2分程度で終わります。英語で話している感覚について回答します。

スピーキングタスクについて
モノローグスピーキングタスクがそれぞれ 2 分かかります。グループディスカッションがそれぞれ 9 分（又は12分）かかります。テストを含めて、クラスアクティビティに似ていますので、練習するための良い機会です。与えられたトピックに関して話します。

> モノローグトピックの例:
> Say as much as you can about your favorite season (activities, weather, fashion, food, trips, etc.)
> ディスカッション質問の例:
> 1. What is the best season in Japan?
> 2. Which do you think is better, summer or winter?

録音や録画は英語のみでタスクを行います。インストラクターはサポートできません。

研究者は、あなたのスピーキングを聞いて、英語が時間の経過とともにどのように変化するかにつていて研究します。調査が終わって、自分の結果を知りたい場合は電子メールで研究者に連絡をしてください。

この研究で私の役割は何ですか？
この研究に参加する場合は、研究の開始時にバックグラウンドアンケートに回答し、モノローグスピーキングタスク,グループディスカッションと別のアンケートを各3回実施します。

この研究に参加していると、私は傷つく可能性があるのでしょうか？
- あなたが英語のスピーキングを録画されることが好きじゃない場合に不快に感じる可能性があります。
- 他の研究者に本当の名前を見せません。さらに録画を見る前、その研究者は研究倫理の研修を行います。個人特定をされるリスクは極めて小さいと考えられます。

この研究に参加するにはお金がかかるのでしょうか？
この調査に参加する費用はかかりません。

この研究に参加していることは、私にとって有益なことなのでしょうか？
私たちは、この研究に参加したあなたに利益を約束することはできません。ただし、英語のスピーキング能力と流暢さを向上させ、クラスアクティビティやテストの練習になります。

研究者は、英語を外国語として学ぶ学生のために英語のスピーキングがどのように発展するかをより深く理解します。より深い理解が、研究者、教師、英語プログラムのコーディネーター、試験作成者などに有益になるかもしれません。

この研究のために収集した情報はどうなるのでしょうか？
個人情報は、この研究を実施または見守る人や組織(該当する場合)と共有される可能性があります:

- この調査の研究者
- この調査の研究者の協力者
- この研究を審査した施設内審査委員会（IRB）
- テンプル大学

この研究に基づく出版や発表が行われる場合があります。しかし、個人が特定されることがないよう、皆さんの個人情報は厳重に取り扱い管理いたします。

完全な秘密を約束することはできませんが、法律で義務付ける範囲内で個人情報を他者に開示することはしません。

この研究に関する私の疑問に答えてくれる人はいますか？
質問、懸念、意見がある場合、またはこの研究があなたを傷つけたり、病気にしたりしたと思う場合は、1ページ目に記載されている上記の電話番号の研究責任者に相談してください。

この研究は、機関審査委員会（IRB）によって監督されています。IRB とは、調査研究の独立した審査を行う人々のグループです。次の項目に該当する場合は、(215) 707-3390 （米国）または irb@temple.edu までご連絡ください。

- 調査チームが回答していない質問や懸念、意見がある。
- 研究チームからの回答が得られていない。
- 研究チームと連絡が取れない。
- 誰かに研究の話をしたい。
- 研究対象者としての権利について質問がある。

私の承諾を得ることなく私がこの研究から削除されることがあるのでしょうか？
この研究の責任者は、あなたの承認なしにこの研究から除外することができます。除外する理由としては：

- バックグラウンドアンケートに回答しない場合。
- 本格的にスピーキングタスクを完了しない場合。例えば、完全にトピック外の回答などです。
- スピーキングタスクを全部完了しない場合。

私たちは、あなたの健康、福祉、またはこの研究に参加する選択に影響を与える可能性のある新しい情報についてお伝えします。

この研究に参加することに同意しても、後で気が変わった場合はどうなりますか？
この研究から辞退される場合は、調査チームに連絡して、調査担当者がご依頼に応じて調査からデータの全部または一部を削除できます。この研究から辞退された後でも、授業でスピーキングタスクに取り組むことができます。研究を辞退されてもペナルティはありません。

本研究協力の同意

私は、本状記載内容及び本研究の実施条件を理解し、同意します。

____

名前（漢字）

______________________________________ ____________________

名前（ローマ字） 日付

APPENDIX C

BACKGROUND QUESTIONNAIRE (ENGLISH VERSION)

Your name: ______________________ Today's Date: ______________

Age: ___________________________ Gender: Male Female Prefer not to answer

Section I: Test scores
List your most recent score for any of the following tests.

Test	Total score/Level	Reading section score	When you took the test (Month/Year)
TOEFL			
TOEIC			
IELTS			
EIKEN			

Section II. Overseas Experience
List the countries you have visited, when you visited them, and the length of your stay in that country stating with the most recent trip overseas.

Country visited (e.g., Canada)	When you visited the country (e.g., 2019)	Length of your stay in the country (e.g., 2 weeks)	Purpose of your visit (e.g., vacation, English study)
1.			
2.			
3.			
4.			
5.			
6.			
7.			
8.			

Section III. English Study Experience—Formal Education
Did you study English in your elementary school? If yes, describe your experience:

Describe your experience with English in junior high school (e.g., conversation tasks, grammar translation tasks)

Describe your experience with English in high school (e.g., conversation tasks, grammar translation tasks)

Were you ever in a school club in which you used English? If yes, describe your experience.

Section IV. English Study Experience—Outside of School
List any experience you have studying English outside of a formal school setting. Typical examples are cram school, English conversation school, and private English lessons.

Study context (e.g., cram school, English conversation school	When did you study? (e.g., 2019)?	How long did you study? (e.g., 1 year)
1.		
2.		
3.		
4.		
5.		
6.		
7.		
8.		

APPENDIX D

BACKGROUND QUESTIONNAIRE (JAPANESE VERSION)

氏名: ________________________________ 日付: ______________________

年齢: ______________________ 性別: 男 女
回答したくない

Section I: テストスコアについて
以下の各テストについて、もっとも最近の合計スコア（英検は何級を取得したか）およびリーディングセクションのスコア、受験した時期を記入して下さい。

テスト	合計スコア/級 (例: 85)	リーディングセクションのスコア (例: 24)	受験した時期 (例：2019 年 3 月)
TOEFL			
TOEIC			
IELTS			
英検			

Section II. 海外滞在経験について
あなたの海外滞在経験について、もっとも最近のものから順に、滞在経験のある国および滞在した時期、滞在期間、滞在目的を以下に記入して下さい。

滞在した国 (例: Canada)	滞在した時期 (例: 2019)	滞在期間 (例: 2 週間)	滞在目的 (例: 休暇、英語学習)
1.			
2.			
3.			
4.			
5.			
6.			
7.			
8.			

Section III. **英語学習に関する経験について — 学校教育**
あなたは小学校の授業で英語を学びましたか。もし学んだ場合、そこでの学習について教えて下さい。

あなたの中学校での英語学習の経験について詳しく教えて下さい。
（例: 会話活動、文法訳読方式の活動）

あなたの高校での英語学習の経験について詳しく教えて下さい。
（例: 会話活動、文法訳読方式の活動）

あなたは今までに学校で英語を使う部活動に所属したことがありますか。もし所属した経験があれば、そこでの活動について詳しく教えて下さい。

Section IV. **英語学習に関する経験について — 公教育外（学校外）**

公教育外（学校外）で、英語を学習した経験について、その教育機関、学習した時期、学習期間を以下に記入して下さい。
※典型例としては、塾や、英会話スクール、個人の英語レッスンなどが挙げられます。

教育機関など (例: 塾、英会話スクール)	学習した時期 (例: 2019)	学習期間 (例: 1 年間)
1.		
2.		
3.		
4.		
5.		
6.		
7.		
8.		

APPENDIX E

LIST OF FORMULAIC PHRASES AND QUESTIONS

Giving Opinions

Asking for opinions	Giving opinions
- What do you think (about...)?	- In my opinion...
- What's your opinion, (name)?	- (Personally speaking), I think...
- What does everyone (else) think?	- I'm not sure, but I think...

Giving Reasons

Asking for reasons	Giving reasons
- Why do you think so?	- It's mainly because...
- Can you tell me why?	- One reason is...
- How come?	- Another reason is...

Agreeing and Disagreeing

Agreeing	Disagreeing
- I agree (with you)	- I partly agree
- I think so too	- I see your point, but...
- I totally / completely agree	- I'm not so sure about that...

Joining a Discussion

Asking others to join	Joining a discussion
- What does everyone think?	- Can I start?
- Who would like to start?	- Can I say something?
	- Can I ask a question?

Giving Examples

Asking for examples	Giving examples
- For example/instance?	- For example/instance, ...
- Such as?	- One example is...
- Can you give me an/another example?	- Another example is...

Follow-up Questions

Who...?	What...?	Which...?	When...?	Where...?	Why...?
How...?	Do you...?	Are you...?	Can you...?	Have you...?	

Talking about Possibilities

Asking about possibilities	Talking about possibilities
- If... what *will* you...?	- If... I *will*...
- If... what *would* you...?	- If... I *would*...

Choosing Topics

Asking which topic to choose	Suggesting a topic
- What shall we discuss first?	- Why don't we discuss *(topic)*?
- What shall we discuss next?	- Let's discuss *(topic)*

Paraphrasing

Paraphrasing someone's comment	Paraphrasing your own comment
- Do you mean...?	- I mean...
- So, in other words...?	- In other words, ...
- Are you saying...?	- What I'm saying is...

APPENDIX F

EXAMPLE DISCUSSION TEST SCORESHEET

Student 1:

Listening	Speaking
<Notes>	
	Total /25
<Communication> Reac-tions ☐☐☐☐☐ ☐☐☐☐☐ Ag&Dis ☐☐☐☐☐ ☐☐☐☐☐	<Content> Speaking turns ☐☐☐☐☐☐☐☐☐☐☐ Total /5
NoM ☐☐☐☐☐ ☐☐☐	☐☐☐☐☐ ☐☐☐ Total /5
<Questions> FQs ☐☐☐☐☐ ☐☐☐☐☐ OQs ☐☐☐☐☐ ☐☐☐☐☐	<Questions> TTQ ☐☐☐☐☐☐ ☐☐☐☐☐☐
	Total /5
<Function> Op ☐☐☐ JD ☐☐☐ Rea ☐☐☐ Ex ☐☐☐	<Function> Op ☐☐☐ JD ☐☐☐ Rea ☐☐☐ Ex ☐☐☐
	Total /5
<Language> Unclear ☐☐☐☐☐☐ Japanese ☐☐☐☐☐☐	Total /5

Student 2:

Listening	Speaking
<Notes>	
	Total /25
<Communication> Reac-tions ☐☐☐☐☐ ☐☐☐☐☐ Ag&Dis ☐☐☐☐☐ ☐☐☐☐☐	<Content> Speaking turns ☐☐☐☐☐☐☐☐☐☐☐ Total /5
NoM ☐☐☐☐☐ ☐☐☐	☐☐☐☐☐ ☐☐☐ Total /5
<Questions> FQs ☐☐☐☐☐ ☐☐☐☐☐ OQs ☐☐☐☐☐ ☐☐☐☐☐	<Questions> TTQ ☐☐☐☐☐☐ ☐☐☐☐☐☐
	Total /5
<Function> Op ☐☐☐ JD ☐☐☐ Rea ☐☐☐ Ex ☐☐☐	<Function> Op ☐☐☐ JD ☐☐☐ Rea ☐☐☐ Ex ☐☐☐
	Total /5
<Language> Unclear ☐☐☐☐☐☐ Japanese ☐☐☐☐☐☐	Total /5

Student 3:

Listening	Speaking
<Notes>	
	Total /25
<Communication> Reac-tions ☐☐☐☐☐ ☐☐☐☐☐ Ag&Dis ☐☐☐☐☐ ☐☐☐☐☐	<Content> Speaking turns ☐☐☐☐☐☐☐☐☐☐☐ Total /5
NoM ☐☐☐☐☐ ☐☐☐	☐☐☐☐☐ ☐☐☐ Total /5
<Questions> FQs ☐☐☐☐☐ ☐☐☐☐☐ OQs ☐☐☐☐☐ ☐☐☐☐☐	<Questions> TTQ ☐☐☐☐☐☐ ☐☐☐☐☐☐
	Total /5
<Function> Op ☐☐☐ JD ☐☐☐ Rea ☐☐☐ Ex ☐☐☐	<Function> Op ☐☐☐ JD ☐☐☐ Rea ☐☐☐ Ex ☐☐☐
	Total /5
<Language> Unclear ☐☐☐☐☐☐ Japanese ☐☐☐☐☐☐	Total /5

Student 4:

Listening	Speaking
<Notes>	
	Total /25
<Communication> Reac-tions ☐☐☐☐☐ ☐☐☐☐☐ Ag&Dis ☐☐☐☐☐ ☐☐☐☐☐	<Content> Speaking turns ☐☐☐☐☐☐☐☐☐☐☐ Total /5
NoM ☐☐☐☐☐ ☐☐☐	☐☐☐☐☐ ☐☐☐ Total /5
<Questions> FQs ☐☐☐☐☐ ☐☐☐☐☐ OQs ☐☐☐☐☐ ☐☐☐☐☐	<Questions> TTQ ☐☐☐☐☐☐ ☐☐☐☐☐☐
	Total /5
<Function> Op ☐☐☐ JD ☐☐☐ Rea ☐☐☐ Ex ☐☐☐	<Function> Op ☐☐☐ JD ☐☐☐ Rea ☐☐☐ Ex ☐☐☐
	Total /5
<Language> Unclear ☐☐☐☐☐☐ Japanese ☐☐☐☐☐☐	Total /5

APPENDIX G

SAMPLE LESSON HANDOUT

Discussion Skill: Agreeing and Disagreeing

Agreeing	Disagreeing
- I agree (with you)	- I partly agree
- I think so too	- I see your point, but...
- I totally / completely agree	- I'm not sure about that...

Be careful...

- Agreeing and disagreeing is a useful way of connecting ideas and comments in a discussion
- It also shows that you are listening and have understood your classmates' comments
- It's often better to disagree indirectly than directly!

Let's practice!

Discuss these questions with your partner. Remember to agree or disagree.

1. Where is the best place for a relaxing vacation?
2. Where is the best place to go on a school trip?
3. Where is the best place to eat delicious food?
4. Where is the best place in Japan for a foreign tourist to visit?

Discussion 1: Country versus City

For each situation below, decide which is the best place to live: the countryside, the suburbs, or the city.

1. Growing up *(country / suburbs / city)*
2. Making new friends *(country / suburbs / city)*
3. Going to high school *(country / suburbs / city)*
4. Finding a good job *(country / suburbs / city)*
5. Relaxing at the weekend *(country / suburbs / city)*
6. Sightseeing *(country / suburbs / city)*
7. Shopping *(country / suburbs / city)*
8. Retiring *(country / suburbs / city)*

Compare your ideas with a partner

Discussion 1

1. Which is the best place to live: the countryside, the suburbs, or the city?
2. Which is better, living in one place forever or living in different places?

Discussion 2: Deciding Where to Live

Which of the following things are most important when deciding where to live? (Choose three)

Living close to nature	
Living somewhere inexpensive	
Living close to school or work	
Living somewhere peaceful	
Living close to public transportation	
Living close to shops and restaurants	
Living somewhere safe	
Living close to family and friends	

Compare your ideas with a partner

Discussion 2

1. What is important when deciding where to live?
2. Where is the best place to live in Japan?
3. If you lived abroad, where would you like to live?

APPENDIX H

EXAMPLE OF RECORDED MONOLOGUE (TIME 3)

I’m going to talk about my high school life. (mm) I joined soccer club’s manager- soccer clubs and I- I did manager. (mm) It is- it was so hard work because the player is- player- there are- there were v- a lot of players so (er) the mana- management of them is- was very hard. (mm) Next about my teachers. I remember- I remember the teacher who managed world history. (mm) It’s because the teacher supported me- supported me when I- I was hard about university entrance exams. (mm) He- he took the class after the school so I- I could study world history a lot and I- when I ta- took a

APPENDIX I

EXCERPT OF RECORDED GROUP DISCUSSION (TIME 3)

Y: Is high school life enjoyable?
I: Can I say something?
Y: [Sure.
R: Sure.]
I: I think my high school life is enjoyable because I made a lot of memories, for example my clubs and event- school event, and yeah, okay. *(laughs)*
Y: Okay, can I start? *(raises hand)*
I: [Okay.
R: Sure.]
Y: Um my school life was enjoyable but it was so challenging because um I made- I could made a lot of friends, but um high school life um so gave me a lot of um task like studying and club, and I was- I had difficulty in studying all of the subjects and my club activity was so hard and so mm some aspects high school life was enjoyable but other aspects it is- it was so- it was not enjoyable, yeah.
I: What club did you?
Y: I joined the tennis club.
I: Ah tennis club?
Y: Ah senior- senior member- senior member were scared for me, for us, so- also the same- same grade member were also scared for the members because they want to do as- they do everything as they want to do, so it was so challenging.
I: It was so hard.
Y: Yeah, that's all.
R: *(raises hand)* Can I say something?
I: [Okay.
Y: Sure.]
R: My answer is yes it is because er I- I didn't belong to any clubs, but er instead I can use the time to play games with friends or to go to the bowling or karaoke. And at third grade er we can- we can er concentrate to studying together, so the class's atmosphere is good, so it is enjoyable.
Y: Ah, I see. Okay.

APPENDIX J

VIDEO RECORDED TOPICS

Topic 1
Monologue topic
Say as much as you can about your favorite season (activities, weather, fashion, food, trips, etc.)

Discussion questions
1. What is the best season in Japan?
2. Which do you think is better, summer or winter?

Topic 2
Monologue topic
Say as much as you can about your favorite place to visit in Japan (things to see, things to do, food, people, atmosphere, etc.)

Discussion questions
1. What is the best place to visit in Japan?
2. Which do you think is better, visiting the sea or the mountains?

Topic 3
Monologue topic
Say as much as you can about your high school life (clubs, lessons, teachers, uniform, classmates, etc.)

Discussion questions
1. Is high school life enjoyable?
2. Which do you think is harder, high school or university?

Instructions Given to Participants
Given before the group discussion

Please discuss these questions in English only. <u>Please discuss them in the way that you feel is best.</u> You can use what you have learned in class, although it is not necessary to do so. Do not stop before you are told to do so.

この質問を英語で話し合ってください。授業で学んだことを使っても使わなくても構いません。<u>思うままに話し合ってください。</u>ただし、合図があるまで続けてください。

APPENDIX K

SPEAKING SELF-EFFICACY QUESTIONNAIRE (ENGLISH VERSION)

Please use the following scale from 1 to 6 to respond to the following statements. Choose the number that best describes how you feel using the descriptions below.

1	2	3	4	5	6
Strongly disagree	Disagree	Moderately disagree	Moderately agree	Agree	Strongly agree

1. I can express my opinion in English.	1	2	3	4	5	6
2. I can use the phrases and skills that I learn in an English speaking class.	1	2	3	4	5	6
3. I can speak English when talking to a partner in pair work activities.	1	2	3	4	5	6
4. I can speak English when taking part in group discussions.	1	2	3	4	5	6
5. I can listen actively to my classmates during English speaking class.	1	2	3	4	5	6
6. I can solve communication problems in English.	1	2	3	4	5	6
7. I can ask questions in English.	1	2	3	4	5	6
8. I can speak English only for a whole lesson.	1	2	3	4	5	6

APPENDIX L

SPEAKING SELF-EFFICACY QUESTIONNAIRE (JAPANESE VERSION)

以下の各項目につき、どの程度当てはまるか自己評価し、１～６の数字で答えてください。なお、１～６の数字については、以下の基準を参考にしてください。

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	ややそう思う	そう思う	強くそう思う

1.	英語で自分の意見を伝えることができる。	1	2	3	4	5	6
2.	英語のスピーキングクラスで学んだ表現やスキルを使うことができる。	1	2	3	4	5	6
3.	ペアワークの時、パートナーに英語で話すことができる。	1	2	3	4	5	6
4.	グループワークの時、グループ内で英語で話すことができる。	1	2	3	4	5	6
5.	英語のスピーキングクラスで積極的に聞くことができる。	1	2	3	4	5	6
6.	英語のスピーキングクラスで、コミュニケーションの問題がある時、英語で解決できる。	1	2	3	4	5	6
7.	英語で質問をすることができる。	1	2	3	4	5	6
8.	英語のスピーキングクラスの間、英語のみで話すことができる。	1	2	3	4	5	6

APPENDIX M

SPEAKING ANXIETY QUESTIONNAIRE (ENGLISH VERSION)

Please use the following scale from 1 to 6 to respond to the following statements. Choose the number that best describes how you feel using the descriptions below.

1	2	3	4	5	6
Strongly disagree	Disagree	Moderately disagree	Moderately agree	Agree	Strongly agree

1.	I feel nervous when speaking English to a partner in pair work activities.	1	2	3	4	5	6
2.	I feel nervous about participating in English group discussions.	1	2	3	4	5	6
3.	I feel nervous because other students speak English better than I do.	1	2	3	4	5	6
4.	I am worried that other students think that my English speaking skills are bad.	1	2	3	4	5	6
5.	I am worried about making mistakes when speaking English.	1	2	3	4	5	6
6.	I feel nervous when someone doesn't understand what I say in English.	1	2	3	4	5	6

APPENDIX N

SPEAKING ANXIETY QUESTIONNAIRE (JAPANESE VERSION)

以下の各項目につき、どの程度当てはまるか自己評価し、１～６の数字で答えてください。なお、１～６の数字については、以下の基準を参考にしてください。

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	ややそう思う	そう思う	強くそう思う

1.	ペアワークの時、パートナーに英語で話すことは緊張する。	1	2	3	4	5	6
2.	グループワークの時、グループ内で英語で話すことは緊張する。	1	2	3	4	5	6
3.	ほかの生徒が自分より英語が話せると思うと緊張する。	1	2	3	4	5	6
4.	ほかの生徒に自分の英会話力がないと思われていないか心配だ。	1	2	3	4	5	6
5.	英語で話すとき、間違えていないか心配だ。	1	2	3	4	5	6
6.	英語で伝えたいことを相手に理解してもらえなかった時、緊張する。	1	2	3	4	5	6

APPENDIX O

EXTRAVERSION QUESTIONNAIRE (ENGLISH VERSION)

Please use the following scale from 1 to 6 to respond to the following statements. Choose the number that best describes how you feel using the descriptions below.

1	2	3	4	5	6
Strongly disagree	Disagree	Moderately disagree	Moderately agree	Agree	Strongly agree

1.	I am the life of the party.	1	2	3	4	5	6
2.	I like to talk to a lot of people.	1	2	3	4	5	6
3.	I start conversations.	1	2	3	4	5	6
4.	I can get along with almost anyone.	1	2	3	4	5	6
5.	I like to be the center of attention.	1	2	3	4	5	6
6.	I talk a lot.	1	2	3	4	5	6
7.	I feel at ease in a large group of people.	1	2	3	4	5	6
8.	I am skilled at dealing with people.	1	2	3	4	5	6

APPENDIX P

EXTRAVERSION QUESTIONNAIRE (JAPANESE VERSION)

以下の各項目につき、どの程度当てはまるか自己評価し、１～６の数字で答えてください。なお、１～６の数字については、以下の基準を参考にしてください。

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	ややそう思う	そう思う	強くそう思う

1. 私は盛り上げ役である。	1	2	3	4	5	6
2. 私はいろいろな人と話すのが好きだ。	1	2	3	4	5	6
3. 私は人に自分から話しかける。	1	2	3	4	5	6
4. 私はほぼ誰とでも仲良くなれる。	1	2	3	4	5	6
5. 私は注目の的になるのが好きだ。	1	2	3	4	5	6
6. 私はよくしゃべる。	1	2	3	4	5	6
7. 私はたくさんの人の中にいると落ち着く。	1	2	3	4	5	6
8. 私は人と関わるのが上手だ。	1	2	3	4	5	6

APPENDIX Q

END OF COURSE QUESTIONNAIRE

Please answer the questions below in either English or Japanese.
下の質問に答えてください。日本語でも英語でも構いません。

1. In what way(s) do you feel you improved the most during this course?
   このコースの中でどのような点で最も上達したと感じるか。

2. Which activities did you find the most useful in this course?
   このコースの中でどのアクティビティが最も役立つと思うか。

3. Which discussion skills, phrases, or questions did you find the most useful in this course?
   このコースの中でどのディスカッションスキルや表現、質問が最も役立つと思うか。

4. Please write any additional comments you would like to make about the course.
   このコースについて他の意見があれば記入してください。

APPENDIX R

ANALYTIC RATING DESCRIPTORS

	Fluency	Communicative Adequacy
5	• Speaks at almost natural speed • Occasional hesitation at appropriate points between clauses	• Contributions are relevant with appropriate amount of detail • Contributions are always easy to understand
4	• Speaks a little below natural speed • Occasional hesitation at various points	• Contributions are relevant with sufficient amount of detail • Contributions generally easy to understand
3	• Speaks a little slowly • Noticeable hesitation at various points	• Contributions are relevant with limited amount of detail • Contributions are comprehensible
2	• Speaks slowly • Frequent or lengthy hesitation at various points	• Contributions are mostly relevant but lacking in detail • Contributions sometimes difficult to understand
1	• Speaks very slowly • Frequent and lengthy hesitation at various points	• Contributions are often irrelevant • Contributions are difficult to understand

APPENDIX S

EXAMPLES OF TOPIC DEVELOPMENT AND SUPPORT FEATURES

Topic Development Moves

Topic Extension (Statement)

I think (.) Okinawa is the best place (.) to visit because er (.) I visit (.) er Okinawa twice? first is (.6) school trip? and the other is: family trip (.7) and (2.2) er:: (1.3) in Okinawa there were a lot- there were museum about World War Two? (.5) so I learned a lot of things? (1.22) in the (.) museum and (1.4) and I learned about World War Two.

Minimal Extension

I: Can you: sports ano- another sports

(3.0)

→ A: Tennis

Topic Extension (Question)

→ R: Is <u>there</u> cherry blossom in Korea?

S: Yes

Turn Management

→ C: Can I start?

S: [Sure

A: <u>Sure</u>]

Topic Management

→ A: So (.) we move on to next question?

N: [Ok

M: Yeah]

Support Moves

Backchannels

M: I like the cooking club and the English club

→ A: Yeah

Agree/Disagree Markers

R: I think (.5) spring (.) is the best.

→ K: Me too

Confirmations of Comprehension

N: … my teacher always give me interesting story and we often do experiments?... yes that makes me so [happy.

→ K: You] (.6) you means you can en<u>joy</u> (.7) your high school life?

N: <u>Yes</u> (.7) how about <u>you</u>.

Repairs

Y: Girls er played ah volleyball er but er I (.) dislike volleyball?

M: Why

Y: er (1.0) hand is er (.5) hand is ((gestures hitting hand))

→ T: Ha- your hand [hurt?

APPENDIX T

TRANSCRIPTION SYMBOLS

[ ]	Points where overlapping talk starts and ends
(.)	Noticeable within-turn pause of less than 0.5 seconds
(.5)	Length of pause (.5 or above within turn; 1.0 or above between turns)
<u>Underline</u>	Heavily stressed word or syllable
,	Falling intonation within a turn
.	Falling intonation at the end of a turn
?	Rising intonation (not necessarily a question)
-	Abrupt cutting off of sound
:	Lengthened vowel sound (extra colons indicate greater lengthening)
(( ))	Non-verbal conduct (e.g., laughter)
( )	Translation of L1
%	Inaudible or unclear
...	Omitted talk

APPENDIX U

EXAMPLE FACETS SPECIFICATION FILE FOR MFRM RSM ANALYSIS

```
Title = Fluency combined data file
Facets = 3                  ; students, raters, topic
Positive = 1                ; students are able, raters and topics are difficult
Noncenter = 1               ; student measures float
Arrange = M                 ; arrange tables in measure descending order
Usort = U                   ; sort residuals by unexpectedness
Inter-rater = 2             ; rater statistics for facet 2
Model = ?,?,?, R5           ; facets are students, raters, topics, on 5 point rating scale
*
Labels =                    ; labels and additional info for each facet follows
1, Student                  ; students are labeled by number only
011-203
*
2, Rater                    ; raters are labeled by number
1, Rater1
2, Rater2
3, Rater3
*
3, Topic                    ; topics counterbalanced between 2 classes
1, Favorite season
2, Favorite place
3, School life
*
4, Score
1 = Fluency                 ; one score was given for fluency (from 2-5)
*
Dvalues = 4,1               ; data begin in column 4; 1 column of data

Data =
011   1   1   2
011   2   1   3
011   3   1   2
…
```

APPENDIX V

EXAMPLE WINSTEPS CONTROL FILE FOR RASCH RSM ANALYSIS

```
&INST
TITLE = Extraversion Questionnaire
Input Data Format (data at the end of this file)
NAME1 = 1              ; column of start of person information
NAMLEN = 3             ; maximum length of person information
PERSON = PERSON ; persons are called ...
ITEM = ITEM            ; items are called ...
ITEM1 = 4              ; column of first item-level response
NI = 8                 ; number of items = test length
XWIDE = 1              ; number of columns per response
DATA =                 ; data after control specifications
Data Scoring
CODES = "12345"        ; valid response codes
*
&END
EXT1: I am the life of the party.
EXT2: I like to talk to a lot of people.
EXT3: I start conversations.
EXT4: I can get along with almost anyone.
EXT5: I like to be the center of attention.
EXT6: I talk a lot.
EXT7: I feel at ease in a large group of people.
EXT8: I am skilled at dealing with people.
END NAMES
S0123232432
S0234333433
S0312222222
…
```