4-2016

A longitudinal study of the development of fluency of novice Japanese learners: Analysis using objective measures

Saori M. Houston
Purdue University

Graduate School Form
30 Updated 12/26/2015

PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By Saori M Houston

Entitled

A LONGITUDINAL STUDY OF THE DEVELOPMENT OF FLUENCY OF NOVICE JAPANESE LEARNERS: ANALYSIS USING OBJECTIVE MEASURES

For the degree of Master of Arts

Is approved by the final examining committee:

Atsushi Fukada
Chair

April Ginther

Colleen Neary-Sundquist

To the best of my knowledge and as understood by the student in the Thesis/Dissertation Agreement, Publication Delay, and Certification Disclaimer (Graduate School Form 32), this thesis/dissertation adheres to the provisions of Purdue University's "Policy of Integrity in Research" and the use of copyright material.

Approved by Major Professor(s): Atsushi Fukada

Approved by: Atsushi Fukada 4/18/2016

Head of the Departmental Graduate Program Date

A LONGITUDINAL STUDY OF THE DEVELOPMENT OF FLUENCY OF NOVICE JAPANESE LEARNERS: ANALYSIS USING OBJECTIVE MEASURES

A Thesis

Submitted to the Faculty

of

Purdue University

by

Saori M Houston

In Partial Fulfillment of the

Requirements for the Degree

of

Master of Arts

May 2016

Purdue University

West Lafayette, Indiana

ACKNOWLEDGEMENTS

I would first like to thank my advisor and committee chair Professor Atsushi Fukada for providing me great opportunities as a graduate student throughout my time at Purdue University.

I would like to show my gratitude to Professor April Ginther and Professor Colleen Neary- Sandquist for serving as committee members for my thesis and giving support. I would like to thank Marina Ishii for allowing me to collect data from her class.

I would also like to thank Sun Young Kim and Minji Kang for encouraging me throughout this whole process and supporting me.

I also like to thank Mayu Miyamoto, Taku Okamoto, Mai Takeuchi and Namiko Uchida for helping me rate the audio data.

Finally, I must express my very profound gratitude to my family in Japan and in Virginia, and my husband for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them.

TABLE OF CONTENTS

Page

Page

Page

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Houston, Saori M. MA, Purdue University, May 2016. A Longitudinal Study of Development of Fluency of Novice Japanese Learners: Analysis Using Objective Measures. Major Professor: Atsushi Fukada.

Fluency has been studied extensively in ESL and EFL mainly to determine what spoken features are characteristics of fluent speech by comparing students who participated in study abroad programs with those who did not. These studies were mainly done with advanced learners of English as a second or foreign language, and there have not been many studies conducted with novice-level learners of foreign languages. Japanese fluency studies are especially in the minority. It is necessary to investigate the characteristics of fluency in Japanese novice-level learners since Japanese shares very little in common with English.

This study investigated the developmental changes in fluency in Japanese foreign language learners (JFL) over the course of one semester using objective measures. Research questions are 1) which objective measures change in relationship to changes in L2 general proficiency throughout a semester, and how do they change, 2) Which objective measures correlate to subjective rating obtained from Japanese instructors?

The participants were 30 students enrolled in Japanese 101. The objective measures were obtained by annotating audio samples with Praat and Syllable Nuclei and by parsing the annotations and calculating measures with Fluency Calculator (Fukada, Hirotani & Matsumoto 2015). The audio data was collected at the beginning and end of the semester with the same set of tasks. Objective measures taken were speed, quantity of speech, pause related measures and several measures concerned with repairs. Accuracy was also measured by the number of AS-units with or without errors.

The results for the first research question suggested that the speed of speech showed steady development from very early stages in the students' language learning process. Silent pause measures indicated that leaners became able to pause at grammatical junctures as the semester went on, but the overall pause ratio did not seem to decrease between the collection points. In addition, it was found that the two tasks used in this study generated very different results. It is not clear which task was better able to gain access to the learners' true fluency, and this should be further investigated in future studies.

Correlation coefficients were calculated to see the relationship between subjective measures and objective measures in order to answer research question 2. The results indicated that speed related measures showed high correlation values indicating that they could be good measures to predict oral proficiency. Mean run length also showed steady correlations to subjective scoring at both the first and second collection points. Pause related measures showed quite different correlation values from the first to second collection points. There were some measures that changed between the collection

points, so it will be necessary to see how the relationship between oral proficiency and the objective measures may change with a wider variety of learners in future studies.

CHAPTER 1. INTRODUCTION

What is fluency?

People often use the term "fluent" to refer to someone who is proficient in a foreign language. Those referred to as fluent may have a high level of proficiency, and their language skills may be very close to those of native speakers. Chambers (1995) noted that this term is often regarded as interchangeable with "proficiency", which can be problematic. Despite its common use in foreign language teaching, the definition of and what people often regard as "fluent" are not well established. Although there is a common understanding that being able to speak fluently is a major goal of both foreign language teaching and the learners themselves, what it means to be a fluent speaker has yet to be discovered, and there is no unified understanding of fluency for language instructors.

There have been numerous studies completed in the past few decades analyzing the fluency of L2 learners. Fillmore (1979) stated that fluency has multidimensional aspects to it and that it ranges from being able to speak with little or no interference from pauses to being able to speak with a creative mind. It seems that his definitions are in part characteristics of native speakers, and foreign language learners may share only some of them.

Lennon (1995) further stated that there are both narrow and broad senses of the term. In the narrow sense fluency is restricted to temporal aspects of speech production, while in the broad sense it means overall proficiency. In second language acquisition and learning, fluency refers simply to speed and other temporal aspects of speech production (Lennon, 1995), unlike the commonly used definition that is interchangeable with proficiency (Chambers, 1997).

Quantitative analysis of fluency

As this research became more common in English as a second language (ESL) and English as a foreign language (EFL), researchers further refined the definition of fluency in speaking from various perspectives. Findings from these studies contributed to second language acquisition studies in several ways. The first is the conceptualization of fluency, which is often perceived as something that relies on the intuitive judgment of a listener. Language instructors use speaking tests and oral interviews as tools for assessing achievement in a course despite the fact that what is perceived as fluent could be very different from one instructor to the next. Consequently, past fluency studies have looked at speech production from the perspective of to what extent a given speaker of a target language is thought of as fluent. Many seemed to agree that speed related aspects of speech, such as speech rate are correlated with what listeners perceive as fluent speech (Freed, 1995; Ginther, Dimova, & Yang, 2010; Iwashita, Brown, McNamara, & O'Hagan, 2008; Kormos & Dénes, 2004; Lennon, 1990; Riggenbach, 1991; Taguchi & Iwasaki, 2008; Towell, Hawkins, & Bazergui, 1996). The investigators used these observable constructs of speech to predict the target learner's proficiency. Many

researchers, however, concluded that a set of fluency measures needs to be used to fully encapsulate fluency, and there is no one measure that can fully grasp the notion of oral proficiency by itself (Iwashita et al., 2008). Researchers seem to be coming to a consensus that speech rate is a key feature to identifying fluent speakers, but these studies are mostly conducted with learners of English as a second or foreign language, and they target advanced level learners that may already have high proficiency to begin with. Conversely, little has been done to investigate what kinds of characteristics are shown in learners of Japanese as a foreign language. English and Japanese are very different languages and share very little in common in terms of linguistic features, such as morphology, syntax, and phonology. Consequently, it is necessary to investigate the characteristics of fluency from the perspective of the Japanese language, as well.

In addition to fluency, researchers have started to analyze speech production from the aspects of what is being referred to as complexity, accuracy and fluency. Skehan (1999) first introduced the concept that when language learners produce an utterance, they have to pay attention to all three of these aspects of speech production. Higher proficiency students may have automatized some aspects of linguistic forms and do not have to pay attention to them during speech production. This ability enables them to achieve fluency in their speech. The dimensions of complexity is analyzed with measures such as the number of words in AS-units (analysis of speech unit[1]) and the number of clauses in AS-units, and accuracy is measured with the number of errors in AS-units (Takiguchi, 2004). When spoken language is analyzed using these three aspects of

[1] Foster et al. (2000)

speech, it is clear that fluency develops in relationship to the other components of speech production. This leads to a better understanding of the developmental path L2 learners undergo while improving their speaking skills.

Fluency as overall proficiency

Subjective scoring is often used to measure overall oral proficiency in order to compare it to objective measures of fluency. The ACTFL Oral proficiency test and their levels were commonly used in the fluency studies mentioned above. However, some researchers pointed out that this tool may not be the best for fluency studies because the way the interview is executed may not reflect learners' fluency at its best (Freed, 1995; Lennon 1990). Furthermore, instructors rarely use the ACTFL OPI as a benchmark when assessing speaking performances in language courses, but rather, use their own rating systems. The inconsistency in spoken language assessment is due to a lack of resources for identifying characteristics of fluent speech and other components that can determine a learner's proficiency level. The use of the communicative teaching method and its focus on speaking skills are crucial in language courses. Subsequently, there needs to be a good indicator of fluency as a measure that can estimate overall proficiency more objectively.

Fluency studies in Japanese L2 learners

Fluency in second language learners has been widely studied in the past decade mainly to conceptualize more objectively what it means to be a fluent speaker. These studies provided a possibility to objectively capture what is normally judged intuitively by listeners. However, these studies focused on ESL and EFL learners, and there are not

many studies of Japanese as a foreign language (JFL) learners. Iwashita (2006) analyzed the spoken language of EFL and JFL students in terms of complexity features, and she found that the EFL and JFL learners followed different developmental paths. This result further suggests that JFL learners' fluency needs to be investigated separately to determine which characteristics are language specific. Despite the claim by many that a large set of measures are needed to assess fluency, JFL and JSL learners' spoken language has been studied in the past with only a relatively small set of measures (Ishizaki, 2004; Ishizaki, 2005; Uchida, 2005). Additionally, these studies focused on advanced level learners, and novice level learners were absent. The speaking skill is not something that is introduced later on in a language course, and its development starts right as learners begin to study a language. For this reason, the present study targets novice learners.

Motivations for the study

The present study addressed the gap identified above by using a large set of measures to analyze how Japanese L2 novice learners' oral proficiency develops. More than 20 measures were used in all. It analyzed speech productions from novice Japanese learners who were enrolled in a Japanese 101 course at Purdue University. Furthermore, the speaking performances of JFL learners were analyzed twice in the semester in order to trace the developmental path they took. This study also investigated how subjective scoring correlates to each of the objective measures in order to understand which measure(s) may best represent proficiency.

Research Questions

The study is exploratory in nature and will be guided by the following research questions.

Research Question 1: Which objective measures change in relationship to changes in L2 general proficiency throughout a semester, and how do they change?

Research Question 2: Which objective measures correlate to subjective rating obtained from Japanese instructors?

CHAPTER 2. LITERATURE REVIEW

The goal of this study is to analyze developmental changes in fluency in Japanese foreign language learners (JFL) over the course of one semester. This study also investigates the relationship between objective measures and subjective scores to see to what extent overall proficiency can be estimated using the selected objective measures. This chapter discusses the general definitions of fluency and those used in second language acquisition (SLA). A review of literature is then given in terms of fluency observed by quantitative analysis, pausing in speech in relationship to fluency, and proficiency levels and fluency. As revealed by the studies mentioned above, the definition of fluency and what determines proficiency are still not agreed upon. Finally, since a major target of fluency studies is English as a second or foreign language, the necessity of conducting a study in Japanese is discussed.

Fluency as a general term referring to L2 proficiency

A major goal of many foreign language learners is becoming able to speak L2 fluently like a native-speaker. People often refer to someone who is proficient at speaking a foreign language using phrases like, “She is fluent in Japanese.” Chambers (1990) said, “In ordinary life, fluency has an extended meaning and is used as a synonym for overall oral proficiency” (p.535). Most people agree that fluent speakers somewhat resemble

native speakers of the target language. Those who sound fluent may be considered so as a result of achieving a high proficiency level or because they are good at imitating characteristics of a native speaker in the target language. Smooth delivery of a message is an important aspect of fluent speech, and it contributes to successful communication. Speech that is interrupted by too many pauses, restarts, or self-corrections may be a hindrance to successful communication in L2.

When subjectively assessing the speech production of L2 learners, pronunciation and grammar play an important role, and instructors assess these factors by their accuracy of the produced forms or pronunciation. Fluency also plays an important role, but unfortunately, language instructors know little of how to assess this aspect of speech, and there are not many studies concerning the fluency of JFL learners, in particular. However, instructors may be able to assess fluency together with forms, pronunciation, and other areas of speech production. As Chambers said, fluency is an interchangeable term with overall proficiency; fluency is key to giving the impression to listeners as to whether or not one can speak well. Is sounding native-like a precise enough description for language instructors to assess the fluency of L2 learners? What exactly is perceived as native-like fluency? Unfortunately, little is known about how learners develop their fluency in tandem with the development of their overall proficiency.

Definitions of fluency

One of the first works on fluency was by Fillmore (1979), in which he defined fluency as a general term that characterized the speech of native speakers. He defined fluency from four different perspectives as follows:

1. Ability to talk at length with few pauses
2. Ability to talk in coherent, reasoned sentences
3. Ability to have appropriate things to say in a wide range of contexts
4. Ability to deliver a message in a "noble way"

Fillmore (1979:51)

These definitions and characteristics of fluent speech cover many aspects of spoken language. They show that fluency involves linguistic knowledge as well as social and creative ability. However, these may not be appropriate descriptions to use when assessing one's foreign language ability, especially for novice learners. Some characteristics of fluency as described by Fillmore may only be applicable to native speakers or highly proficient learners. There is one aspect that can be applied to second language learners, however, which is "the ability to talk at length with few pauses." This notion of fluency has been widely used in second language studies and has been redefined by many language researchers. Lennon (1990), for example, identified two senses of the term: broad and narrow. In the narrow sense, fluency is restricted to the temporal aspect of speech production, while in the broad sense it regards fluency as overall proficiency in L2. Skehan (2009) further breaks down Lennon's narrow sense of fluency into three subcategories: (1) breakdown, (2) repair and (3) speed fluency. Breakdown fluency refers to the silent and filled pause phenomena in speech, repair fluency refers to repetition or false starts, and speed fluency refers to speed of one's speech. These three areas have been operationalized in order to analyze both fluency and overall proficiency more objectively. Breakdown fluency is often measured using the number of filled and silent pauses, pause ratio, and pause length. Repair fluency includes

the number of self-corrections and restarts. Speed fluency is found to be a good indicator of both fluency and overall proficiency. They are often operationalized as articulation rate, the total number of syllables per minute, mean length of run, and speech rate, the number of meaningful syllables per minute without filled pauses.

Quantitative approach to fluency

Many researchers focused on the temporal aspect of fluency in order to analyze it reliably. Lennon (1990) investigated four participants in Germany learning English as a foreign language to determine their fluency development before and after a study abroad program. He found that speech rate increased between 13 and 30 percent while the number of filled pauses decreased between 22 and 60 percent over 21 weeks. Other studies also agree that speech rate is considered one of the most promising measures correlated to subjective scoring (Ginther et al., 2010; Iwashita, 2010; Kormos & Dénes, 2004, Riggenbach, 1991; Taguchi & Iwasaki, 2008), and it is believed to be one of the noticeable characteristics of fluent speech. Other fluency features that seem to be correlated to perceived fluency include filled pauses and pause time.

Freed (1995) conducted a similar study on a group of L2 French students who participated in a semester-long, approximately sixteen-week study abroad program and those who stayed on campus in U.S. university. Fifteen participants from each group were investigated. ACTFL's Oral Proficiency Interview was used to assess the speech of all the participants, and fluency measures, such as the number of pauses and speech rate, were used to compare the groups. The result was that students who studied abroad showed a faster rate of speech, ranging from 93.5 to 143 words per minute, than on-

campus students, ranging from 71 to 104 words per minute. Additionally, fewer dysfluency factors, such as dysfluency clusters per 100 words, were found in the study abroad group, which ranged from 2.3 to 6.7. Dysfluency clusters per 100 words for the on-campus students, on the other hand, ranged from 2.9 to 8.5. In her study, Freed found the rate of speech to be a good discriminator between the on-campus and study abroad groups.

Towell et al. (1996) also compared the speech of L2 French students before and after a 6 months' study abroad program where French was the target language. The data were collected in a second year course, before study abroad, and in a third year course, after return from the study abroad program. They found that after the study abroad, articulation rate and mean length of run for the L2 students changed. Speech rate before studying abroad was $M = 136.61$, $SD = 32.09$ and after study abroad was $M = 156.88$, $SD = 28.10$. Articulation rate before studying aboard was $M = 385$, $SD = .43$ and after studying aboard was $M = 417$, $SD = 44$. They concluded that the primary factor for fluency development is an increase in length of run and not due to a decrease in pause time or an increase in speed.

The studies reviewed above agree that fluency measures such as rate of speech, articulation rate, and mean length of run are correlated with holistic scales provided by human raters. Other measures found to be possible indicators of fluency were the number of unfilled pauses, total pause time (Iwashita et al., 2008), and pause ratio (Taguchi & Iwasaki, 2008).

The difference in the measures that was found to be a good indicator could be the results of different L2 languages examined. Or it may be because of inconsistency in the

choice of tasks used to elicit speech such as role-play and monologic tasks. Another reason could be inconsistencies in the level of learners that were examined.

Nevertheless, all of them seem to agree that in order to evaluate and detect fluency development precisely, more combinations of measures need to be analyzed and more levels needs to be examined. It is also clear that in order to find a measure that will best indicate changes in oral production in Japanese, the first step is to analyze the oral production of not only advanced level learners but of novice learners as well. By starting with the extremes, we can establish the broad range and then place students within.

Fluency and pauses

The characteristics of pausing largely depend on each individual as to how long and how often they pause. Pauses may be used to plan speech production or to recall correct forms. Zellner (1994) said that a "pause reflects the time needed for the cognitive planning process to catch up." (p47) While native speakers tend to pause at grammatical junctures and rarely use pauses within phrases, L2 leaners often use pauses in these locations perhaps in order to access their linguistic knowledge, which otherwise cannot be retrieved at the same time as they articulate. Pauses longer than 250ms are considered to be associated with constructing a message and are noticeable to a listener (Zellner, 1994). In fact, many fluency studies have used a cutoff time of 250ms for this reason (De Jong, Groenhout, Schoonen, & Hulstijn, 2013; Ginther et al., 2010).

Filled pauses "occur more frequently than false starts and repeats, but not as often as silent pauses" (O'Connell & Kowal, 2008), and extensive use of pauses and filled pauses within words or grammatical sequences may be an indication of dysfluency.

In this case, the location of the pause plays an important role in how speakers process the utterances they produce. While the location of pauses may be an indication of efficiency in speech production, the length of a pause may largely depend on each individual speaker, and each language may have some distinctive characteristics as to how long people usually pause. Pauses in speech, both silent and filled, provide many clues about fluency development since they are a reflection of cognitive load when L2 speakers formulate their speech (Zellner, 1994; Segalowitz 2010).

Since silent pauses are fundamental to articulating a long sentence, it is difficult to distinguish when silent pauses should be considered dysfluency or hesitation phenomena. Filled pauses, in particular, are an indicator of how complex a sentence the speaker is trying to make. Watanabe, Den, Hirose, & Minematsu (2004) suggested that as speech becomes more complex, the ratio of filled pauses also increases. They measured the filled pauses that occurred before dependent and independent clauses, the filled pauses that occurred after topic and case markers, and assigned different levels of complexity based on the types of clause and case marker. Their results showed that the more complex the following clause type is the more filled pauses were found. Their results also suggest that filled pauses may be used as a proxy for the cognitive levels of learners in retrieving linguistic knowledge and their ability to produce speech efficiently since filled pauses were found followed by the most difficult clause type. Thus the location of pauses plays an important role in judging learners' speaking skills. Fluency studies in Japanese are often concerned exclusively with the pausing phenomenon (Ishizaki, 2004; Ishizaki, 2005). These studies analyzed mainly how L2 Japanese learners used silent pauses when

they speak, and the results indicate that learners tend to pause at unnatural locations, which is believed to be a reason that listeners may find their speech difficult to parse.

Complexity, accuracy and fluency

Skehan (1999) recognized that there were three main aspects comprising the production of speech: 1) fluency, 2) accuracy and 3) complexity. When producing an utterance, fluent speakers are able to control all of these equally in order to make a smooth speech delivery. In the case of advanced and native speakers, many aspects of speech become automatized. As a result, they need not pay conscious attention to production. For example, an advanced learner who has acquired perfect grammar and pronunciation does not need to put conscious effort into producing accurate speech, and he or she can focus more on the complexity component in speech production. Fluency here is seen as a sign of how these demands are met when L2 speakers speak, with pause phenomena and repair, in particular, revealing how these attentional demands may affect production. It is then reasonable to assume that fluency can be a very good indicator of developing L2 proficiency and provide beneficial information for instructors to use in promoting fluency as needed.

Fluency observed at a variety of proficiency levels

Many studies in the past examined the fluency of advanced learners, such as those of Towell (1996), Freed (1995) and Lennon (1990). They found differences in the advanced level learners regarding temporal aspects of fluency that are noticeable to listeners, but their studies did not target novice level learners who had just started learning a foreign language. This could be because fluency is seen as a token of achieved

proficiency. It is beneficial to know how objective measures can distinguish between fluent and non-fluent speakers, but fluency also needs to be analyzed with respect to how it develops along with language learning. As Skehan (1999) stated, not only is fluency development important, it is necessary for learners to acquire all three dimensions of speech production, fluency, accuracy, and complexity, in order to smoothly produce the foreign language.

Takiguchi (2004) observed Japanese EFL learners in junior high school at three different times. He concluded that at the beginning stage of learning, participants develop their speed and length of production with accuracy and complexity coming later on. Fluency, on the other hand, showed steady growth throughout.

Fujimori and Koizumi (2011) conducted a similar study with university students and high school students. Their study found that at the low level (high school), students' accuracy did not develop while fluency and complexity did develop. Larsen-Freeman (2006) conducted a study which observed five EFL learners and their development of accuracy, complexity, and fluency in writing, and she concluded that each of the participants followed different paths as to how these three dimensions of production progressed.

Fluency studies like these give insights into how learners develop fluency in relationship to their overall proficiency. In addition, while most fluency studies mainly focused on students in a study abroad condition (Freed, 1995; Lennon, 1990; Towell, Hawkins, & Bazergui, 1996), Takiguchi (2004) and Fujimori and Koizumi (2011) show that learners can develop their fluency in a classroom setting. This is particularly encouraging for instructors as they seek to create activities that focus on fluency. For

fluency studies in English, there are well established sets of measures that most studies use. For example, it is common for speech rate, articulation rate, pause ratio, number of silent and filled pause to be used in these fluency studies. English and Japanese, however, share very little in common in terms of lexicon, morphology, or syntax. It is therefore necessary to first investigate Japanese using a wide variety of measures. As can be seen from these findings, a relatively large set of measures have been used in second language studies (Takiguchi, 9 measures, Fujimori and Koizumi, 17 measures), but most of these studies either focused only on intermediate to advanced level learners, used a very small sample size (about 15 participants in each level were examined), or analyzed only written production from L2 learners. Most importantly, there have only been a few studies concerned with learners of Japanese as a foreign language. It is worth investigating measures that detect changes in accuracy and complexity in addition to fluency in order to fully grasp how learners progress from the very beginning stage of their learning process.

Negishi (2012) investigated 135 EFL students in Japan at three educational levels using proficiency ratings that modeled Common European Framework of Reference for Languages (CEFR: Council of Europe, 2001) to see if they correlated to fluency measures. The measures that showed differences between the education levels were the total number of syllables and words, length of pauses, and total speaking time including pauses. Proficiency ratings derived from CEFR were correlated to the total number of syllables and words. Negishi (2012) concluded that the proficiency score did not seem to correlate to the rate at which the students spoke, but this may be due to the low proficiency level of the students. She mentioned that in other studies, such as by Kormos

and Denes (2004), more advanced students were used to measure fluency, and novice level speakers clearly showed different results regarding the correlation between proficiency and fluency.

Assessment of overall proficiency

Fluency studies so far have used proficiency levels to distinguish the characteristics that exist between learners in terms of temporal fluency measures and other objective measures. In most cases, proficiency level was determined according to ACTFL Oral Proficiency Interview scale (Baker-Smemoe, Dewey, Bown, & Martinsen, 2014; Freed, 1995; Segalowitz & Freed, 2004; Taguchi & Iwasaki, 2008). Oral interview tests given in foreign language courses rarely use the ACTFL OPI to rate learners' achievement and rather employ their own rating scale for oral proficiency. Despite the fact that some researchers claim that the ACTFL OPI may not be the best tool to use to compare against objective measures (Freed, 1995; Lennon 1990), the majority of fluency studies use the ACTFL OPI rating or education level and length of time in a language course as a proxy for proficiency levels. It is often the case that language courses still place more emphasis on measuring grammar points and vocabulary using written tests and not oral tests due to the difficulty of developing a rating scale that best captures learners' ability. Being able to assess spoken language is becoming even more crucial as foreign language teaching is leaning toward a focus on the development of speaking skills. In addition to written tests that assess accuracy in grammar points and vocabulary, being able to assess speaking performance with reliable measures that can capture language achievement and are easily adaptable is increasingly crucial to language

teaching, and a lack of resources necessary to determine what exactly fluent speech is may result in not enough time devoted to it for language teaching in the classroom. The more instructors know about how fluency and speech production develops, the more instructional methods can be applied to their teaching.

Research gaps and motivation for the present study

Fluency studies in ESL and EFL had established some potential that objective measures of fluency can be used to predict proficiency of one's L2 speaking (Ginther et al., 2010; Iwashita et al., 2008). It is widely accepted that measures related to speed, speech rate, and pausing are good indicators of fluency, as well as overall proficiency but specific range of objective measures per proficiency are yet to be found. There are some studies (Iwashita et al., 2008; Lennon, 1990) in which researchers have claimed a combination of measures need to be used in order to determine one's proficiency level more accurately, and no single measure can take the place of a human holistic ratings. Years of research concerning the temporal aspects of fluency have contributed findings in ESL and EFL settings, but we are far from being able to establish a reliable set of measures for Japanese. Several studies have been conducted on the pausing pattern of L2 Japanese learners with different native languages (Ishizaki, 2005; Uchida, 2005), and listenability and pausing effects have also been studied (Ishizaki, 2004), but comprehensive studies using objective measures to look at the other aspects of speech production are still uncommon in Japanese. Unlike ESL and EFL learners, little has been done to analyze the fluency of L2 Japanese learners, and it is difficult to make inferences from research on a language such as English that is far removed from Japanese in terms

of morphology, syntax, and lexicon. For example, Iwashita (2006) examined complexity measures between EFL and JFL students, and the findings from these two L2 groups were quite different. It is clear that regardless of the abundance of studies in English, these findings cannot necessarily be applied to fluency in Japanese. Therefore, the purpose of this study is to analyze Japanese L2 learners using a variety of objective measures in order to find a specific set of measures that can be used as a guide to determining proficiency levels, as well as determine the Japanese specific characteristics of L2 Japanese speakers' speech.

Aside from L2 based developmental variations, how differently various proficiency levels of Japanese learners affect objective measures of fluency, complexity or accuracy has not been well studied, either, though some studies have looked at Japanese L2 learners together with several other languages. One such study was conducted by Baker-Smemoe, Dewey, Bown & Martinsen (2014). They investigated how fluency measures differ across different ACTFL OPI proficiency sublevels without regard to the five L2 languages used. They examined Japanese, German, Russian, French and Arabic learners who were native English speakers. The objective measures of fluency used in this study were syllables per second, number of runs, length of runs, number of pauses, length of pauses, number of hesitations, and number of false starts. Their descriptive statistics showed that clear distinctions by fluency measures in proficiency sublevels were seen in advanced level learners, but there were no clear differences in lower proficiency levels. This result can be interpreted that the measures they used were not capable of detecting changes in the characteristics of novice level learners, and they needed to examine other fluency measures in order to capture these changes. The result

of this study come from mostly European language learners since there were only 4 advanced level L2 Japanese learners out of 59 participants in advanced level learners. Additionally, the only novice level learners in their study were Japanese L2 speakers (*N* = 5 out of 16 Japanese L2 learners), further skewing the study toward advanced level language learners as opposed to the novice level. The findings from this study are not clear as to whether these were language specific or universal characteristics since the level distribution of each language is not well balanced. However, the results suggest that distinguishing the differences among novice level learners is much more complicated than for advanced level learners and may require a larger set of fluency measures to discriminate. This may be because the Japanese language participants were at a relatively lower level than the participants for the other languages who were intermediate to advanced in ability. The distinctive characteristics found for Japanese may not be a result of the language itself but more of an effect of the levels that the Japanese learners were at (novice to intermediate). Consequently, there may be differences in objective measurements depending on the level of learners, and it could also be the case that Japanese is dissimilar enough that there are many aspects that are not comparable to European languages.

What is known from studies targeting ESL and EFL learners is that fluency, such as speech rate and articulation rate can be used to predict proficiency levels together with several other objective measures (Ginther, Dimova, & Yang, 2010). However, how fluency develops in the long term in a classroom setting while looking at the correlation between subjective and objective scoring has not been examined, with the exception of a few studies that targeted EFL learners (Fujimori & Koizumi, 2011; Takiguchi, 2004).

Small sets of measures, such as pausing pattern, have been studied in order to see their relationship to how they are perceived by listeners with high levels of Japanese ability (Ishizaki, 2004; Ishizaki, 2005; Uchida, 2005). Sakuragi (2011) looked at 113 Japanese L2 learners at the intermediate to advanced levels using 10 variables related to fluency, accuracy and complexity. This is one of the few studies that have specifically targeted Japanese L2 learners, but novice level learners are absent from this study, and the set of measures used is relatively small. As Negishi (2012) pointed out, some variables are not compatible with lower level students, and this is one of the problems with previous studies that looked at intermediate to advanced level learners. Together with pauses, Japanese L2 learners' speech production should be analyzed more extensively with a wide variety of measures longitudinally from early in the language learning process, in order to examine the overlooked areas of developmental changes in the learners.

The objectives of the present study

In order to fill the gaps in the literature identified above, the focus of this study is to investigate the fluency of learners of Japanese as a foreign language using various objective measures. This study also examines whether or not proficiency ratings given by Japanese instructors are correlated to these objective measures, and if so, how. Given that many past studies were only concerned with more advanced learners of the target languages, this study investigated JFL learners from the beginning of their language learning process. The development of JFL speech performance was analyzed by collecting speech samples twice in a semester. The next chapter explains the design of the study in detail.

CHAPTER 3. METHODOLOGY

Overview

This study analyzed how L2 Japanese learners developed their fluency over one semester in a college classroom setting by using objective measures. The study was designed to collect data at both the beginning and end of the semester to measure fluency development in 30 students from a first year Japanese course (Japanese 101). This study also investigated which objective measures correlated to the subjective proficiency rating obtained from Japanese instructors. This is to examine the possibility of estimating proficiency with objective measurements. The audio samples came from students who took Japanese 101 in the fall of 2015 at Purdue University. In this course, an online oral practice program called *Speak Everywhere* (Fukada, 2013) was used extensively, and chapter tests were given through this online program. Speech samples from two of the chapter tests that included identical monologic tasks were used to investigate the development of fluency and overall proficiency. There was a span of three months between the two sample collection points. The speech analysis software *Praat* (Boesma & Weenink, 2016) and *Fluency Calculator* (Fukada et al., 2015) were used to obtain measurements from each audio sample.

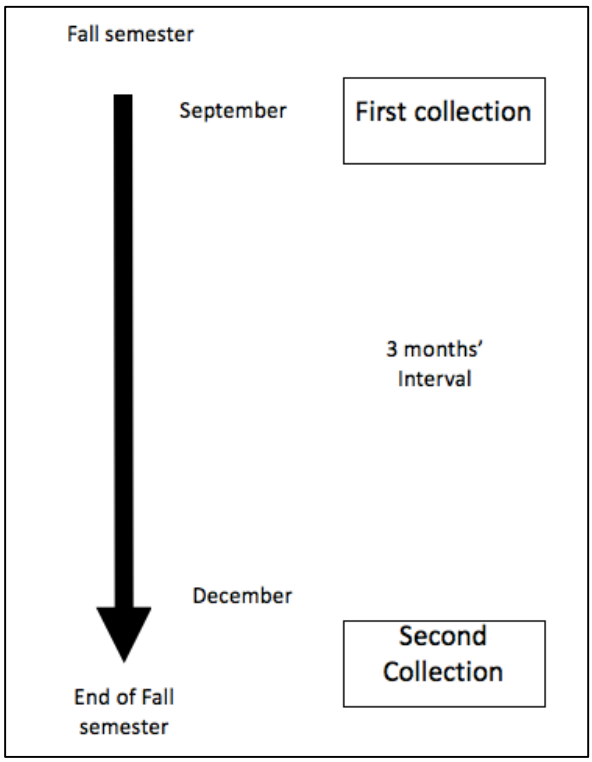

Figure 1. Overview of the study

Participants and Raters

Participants were students enrolled in Japanese 101 in the fall semester of 2015 at Purdue University. Samples were collected from a total of 30 students. The students in this study were enrolled in four different sections of Japanese 101 taught by four graduate teaching assistants along with one instructor who supervised them. Eighteen students that also enrolled in Japanese 102 in the subsequent semester were selected first, and then, another 12 students were added from four sections of the Japanese 101 course. The participants consisted of 10 males and 20 females between the ages of 18-24. Their native languages were English (N = 7), Chinese (N = 20), and Korean (N = 3). Participants'

majors were shown in Table 1. One participant did not have his/her major decided, so the total number participants in the table is 29.

Table 1. Participants' majors

Major	N	Major	N
Electrical Engineering	6	Actuarial Science	1
Computer science	4	Film	1
Agriculture	2	Engineer	1
Fine Art	2	Communication	1
Finance	1	Management	1
Genetics	1	Computer graphic technology	1
Applied exercise	1	Food science	1
Hospitality	1	Accounting	1
Pharmaceutical Sciences	1	Science	1
Psychology	1		

There were four participants who had studied Japanese, and the years of study ranged from 1 year to 4 years (M = 2.5, SD = 1.3). Since a majority of participants were L2 speaker of English (73%, L1 Korean and L1 Chinese), many of them had studied foreign languages before, mainly English (M = 10.5 years, SD = 3.2), Spanish (N = 3, M = 3.3, SD = 1.5), and Chinese (N = 1, 3 years).

The raters in this study were three native Japanese speakers who were either graduate teaching assistants or instructors at Purdue University. Their teaching experience varied from two to nine years (M = 4.8, SD = 3.6) in post-secondary

institutions in the United States. The raters were given a detailed flow-charted scoring system (Miyamoto, 2016) and rater training with sample audio clips. They did not receive a lecture on fluency.

Data Collection

There were six chapter tests given in this semester, and the monologues from two of them were used to investigate fluency development. Since the goal of the study was to analyze the development of oral fluency, two sets of samples were taken three months apart from each other. The online speaking practice software called *Speak Everywhere* (Fukada, 2013) was used to collect the audio. This program was used extensively in Japanese 101, in everything from vocabulary exercises to dialog role-play. The first set of data came from the third chapter test that was given in week seven of the course, and the second set of data came from the sixteenth week of the course.

Task/Material

This course used performance-based oral tests where students were evaluated not only on their grammatical accuracy but also on how naturally and smoothly they gave their responses. To encourage oral practice, practice tests were uploaded at the beginning of each chapter, and the students were encouraged to practice them in order to achieve a fluent and grammatically accurate delivery. This study used two monologic tasks from the chapter tests: (1) give a self- introduction (Self Task), and (2) describe a typical schooldays (School Task). These two tasks were employed at both data collection points in order to achieve comparability. In the first task, subjects were given 120 seconds to give a self-introduction. Self-introduction was one of the first tasks to be introduced in

Japanese 101. In the textbook*, Nakama 1: Japanese Communication, Culture, Context*, self-introduction is taught in Chapter 2 along with a set structures such as “I am from~”, “I am a junior in college”, and “My major is~.” This task was chosen because it is one of the most frequent tasks that learners encounter outside the classroom in real life communication. Although some specific phrases are suggested for use in class, a self-introduction can be performed in various ways from bare minimum to very elaborate depending on one’s level. The tasks to be collected as a speech sample were designed to be simple so that learners can practice them from the beginning of the Japanese learning process and start using them in daily life. Moreover, the task flexibility enables learners to elaborate on their responses in various ways reflective of their abilities as they gain more proficiency.

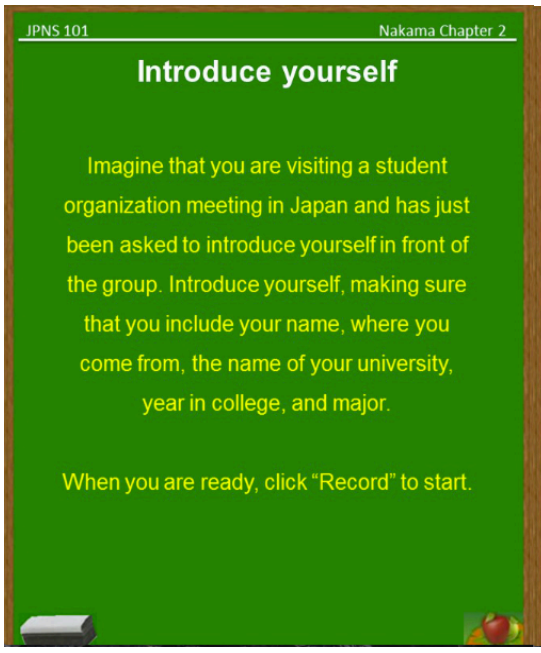

Figure 2. Task 1 Self introduction

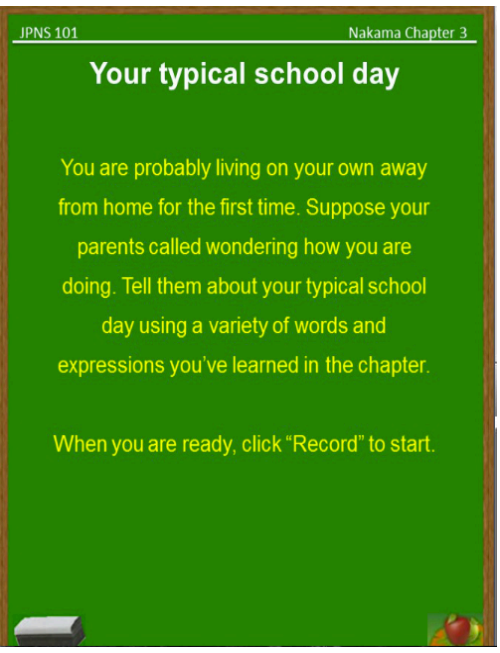

Figure 3. Task 2 Describe your school day

The second task of describing a typical school day routine is also one of the first tasks that were introduced in Japanese 101. This task is also simple, yet the situation of having to describe a day happens more frequently than some other complicated tasks. It is important that the situation where students describe their days can be said in various ways as their Japanese level increases and they become familiar with more complex grammar. This is based on the assumption that when leaners become more advanced, they can use a larger variety of grammar patterns to describe their school day routines. Overall, these tasks were chosen because they are likely to occur in real life outside the classroom, more so than describing a building, for example.

Procedure

All participants in this study took a chapter test every two to three weeks for the total of six times in this course. In Week 6, the Chapter 2 test that contained the self-introduction task was assigned. In Week 9, the Chapter 3 test that contained the second task, describing a school day was assigned. Students submitted their chapter tests within the same day on online speaking exercise website *Speak Everywhere*. In week 17, all of the participants took a final exam that contained the same two tasks. The sample audio collected were downloaded to analyze further in Spring semester of 2016.

Measures

Subjective rating was used as a measure of overall proficiency. A rating system developed by Miyamoto (2016) was used to rate the audio data consistently across all raters. The raters were given training sessions to become familiar with the rating

procedure and to achieve consistency among them. A total of two graduate teaching assistants and a lecturer were recruited to rate the general oral proficiency of the audio samples in terms of "message delivery", "speech delivery", "content", "grammatical accuracy", "speech length and complexity" and "vocabulary and grammar structure." (See Appendix)

Objective measurements were used to analyze the fluency related aspects of the speech samples. Syllable count and silent pauses were annotated by the "Syllable Nuclei" Praat script written by deJong, N. H., & Wempe, T. (2009). Subsequently, the annotations were manually edited to accurately represent syllables, filled pauses, sentences and AS-units (with or without errors). Filled pauses are where the speaker uses fillers such as "mm", "uhh" and "eh". An AS-unit is an "independent clause or sub-clausal unit." (Foster, Tonkyn, & Wigglesworth, 2000) Dysfluency features were also manually coded. These include repetitions, stuttering, and self-corrections. Repetitions are where speakers repeat words or phrases immediately after saying them. Stuttering is also an important feature to fluency as this hesitation phenomena can be seen as a "reflection of the underlying cognitive processes."(Brigitte, 1994) The annotated audio samples were then submitted to Fluency Calculator, which computed fluency related measurements. The objective measurements calculated by Fluency Calculator are listed below.

Table 2. Objective measures

Speech quantity	Total response time.	The time in seconds from the beginning of an audio response to the end
	Total number of syllables.	All syllables in the file
	Effective syllable count.	(Total number of syllables) – (syllables in repeat, stutter, and self-correction intervals)
	Number of Sentences	
Speed	Speech rate.	(Total number of syllables) / (Total response time) * 60
	Articulation rate.	(Total number of syllables) / (Speech time + Filled pause time) *60
	Mean length run.	(Total number of syllables) / (Number of runs) where a run is a sounding interval
	AS-Unit speech rate.	Effective syllable count / AS-Unit time * 60
Pause	Silent pause ratio.	Silent pause time as a ratio of total response time.
	Silent pause count.	The number of all silent pauses longer than 250ms
	Silent pause time.	The sum of time in seconds of the duration of all silent pauses
	Filled pause count.	The number of all filled pauses
	Filled pause time.	The sum of time in seconds of the duration of all filled pauses
	Silent pause count within AS.	The number of silent pauses within AS-Unit intervals
	Silent pause time within AS	The sum of time in seconds of the duration of silent pauses falling within AS-Units
	Silent pause count between AS.	The number of silent pauses between AS-Unit intervals
	Silent pause time between AS.	The sum of time in seconds of the duration of silent pauses falling outside AS-Units
	Filled pause count within AS.	The number of filled pauses within AS-Unit intervals
	Filled pause time within AS	The sum of time in seconds of the duration of filled pauses falling within AS-Units
	Filled pause count between AS.	The number of filled pauses between AS-Unit intervals
	Filled pause time between AS.	The sum of time in seconds of the duration of filled pauses falling outside AS-Units

Table 2. Continued

AS-unit related measures	Number of AS-Units.	
	Number of error-free AS-Units	
	AS-Unit time.	The sum of time in seconds of the duration of all AS-Unit
Repair fluency	Repeat count.	The number of repeats
	Stutter count.	The number of stutters
	Self-correction count.	The number of self-corrections tier

Data Processing

Each audio sample was first recorded in .mp3 format using Camtasia. It was then normalized and noise reduced in order to maximize the performance of Syllable Nuclei (De Jong & Wempe, 2009), which is a Praat (Boersma & Weenink, 2007) script that automatically detects syllables, speech portions, and silent portions in the audio data. The silent time before and after the speech was also eliminated in order to accurately calculate speech time. These procedures were performed using Audacity, a freeware audio editor. The next step was the automated annotation of the audio data using Praat and Syllable Nuclei. Once the automated annotation was completed, the syllables in the audio recordings were manually corrected. Silent and filled pauses, as well as AS-unit (Foster, Tonkyn &Wigglesworth, 2000) boundaries with and without errors, and sound boundaries with repairs were annotated. Pause duration detection was set to 0.25 seconds in this study. This pause duration is a widely recognized standard (Zellner, 1994). All annotated data was saved to a data file with a participant number as its file name in one of two sample collection folders: first collection point and the second collection point. Each folder was submitted to *Fluency Calculator* (Fukada et al., 2015) for processing.

Data analysis

In order to answer Research Question 1, a *t*-test was performed to examine whether or not there were significant differences in the objective measures between the first and second samples. In order to answer Research Question 2, Spearman's rank-order correlation coefficient was computed to examine the correlation between the subjective rating and each objective measure of fluency. Because *t*-test and the significance test on correlation coefficients were performed multiple times, the Bonferroni Correction was applied to avoid type I errors.

CHAPTER 4. RESULTS AND DISCUSSIONS

In this chapter, the collected speech samples are analyzed and discussed. A total of 30 L2 Japanese learners participated in this study. Each learner was given the same two tasks at the beginning and at the end of the semester for a total of 120 data samples. These data were analyzed in order to answer the research questions.

Research Question 1

There were two tasks used in this study to collect data. The first task was a self-introduction (Self task), a monologic task, and the other was a task to describe a school day (School task). The two tasks were introduced in the second and third chapters of the course, respectively, as one of the new tasks to be tested on along with new grammar forms to be used in the tasks. This section examines the results of these tasks and the differences between them in order to answer Research Question 1: Which objective measures develop in relationship to changes in L2 general proficiency in a semester, and how do they develop?

Speed fluency and speech quantity

A summary of the means and standard deviations of speed fluency and speech quantity related measures, along with *t*-test results, are shown in Table 3. The difference

between effective syllable count and the total number of syllables is that the latter includes syllables that fall within the repair portions of speech. As shown in Table 3, there were not large differences between these two measures, indicating that repair did not occur much in this study. However, this study found that the increase in both the total number of syllables and effective syllable count were slightly greater in the Self task than in the School task by *t*-test analysis. The standard deviation for the first data collection (*SD* = 11.59 in the Self task, *SD* = 41.22 in the School task) was smaller than for the second data collection (*SD* = 41.22 in the Self task, *SD* = 61.92 in the School task). Judging from the difference in standard deviations in these two tasks, there were wider variations in speech length in the second collection data for both tasks. The syllable count for these measures includes filled pauses, but as is discussed later in this chapter, there were a very small number of filled pauses. Therefore, increases in syllable count for this study mean that the meaningful speaking time increased.

Table 3. Descriptive statistics of speed fluency and speech quantity.

		First Cillection				Second Collection				
	Task	M	SD	MAX	MIN	M	SD	MAX	MIN	diff
Total number of	i	62.83	11.59	93.00	44.00	81.13	41.22	279.00	50.00	-18.30
syllables	s	165.50	58.75	265.00	71.00	186.97	61.92	301.00	73.00	-21.47
Efective syllable	i	62.57	11.16	89.00	44.00	80.87	41.35	279.00	50.00	-18.30
count	s	162.87	57.72	265.00	71.00	186.00	61.99	301.00	73.00	-23.14
Mean run length	i	10.08	2.70	18.33	5.22	11.26	2.70	18.33	6.09	- 1.18
	s	9.05	2.95	13.88	3.79	11.70	5.20	30.33	4.33	- 2.65 *
Articulation rate	i	349.95	55.03	487.77	229.16	363.70	44.60	467.15	285.88	-13.75
	s	300.36	46.57	419.05	202.46	322.26	41.79	419.25	231.59	-21.91
Number of AS-Units	i	5.13	0.78	10.00	4.00	6.53	2.58	19.00	4.00	- 1.40
	s	9.90	3.07	17.00	5.00	10.03	3.47	17.00	3.00	- 0.14
AS-Unit speech rate	i	333.78	62.64	487.77	157.20	351.29	48.47	467.15	244.60	-17.52
	s	257.74	60.54	375.30	139.81	284.79	58.63	389.09	176.99	-27.05

*p<.05 **p<.01
i= Self introduction , s= School day

Mean run length showed interesting results in that they were almost the same at the second collection (11.26 for Self and 11.70 for School task), for both the Self and School tasks. This was the only measure that showed very similar results between the Self and School tasks, which indicates that mean run length may be a good indicator of a learner's level with the added benefit that it is not affected by task choice. Articulation rate was different between the first and second collections for both the Self and School tasks. When comparing the two tasks, the change was almost the same (1st collection *M* = 300, 2nd collection *M* = 322, +7% in the Self task, 1st collection *M* = 349, 2nd collection *M* = 363, +5% in the School task). The ease of using set phrases can be seen in the resultant faster articulation rate. Learners had an easier time formulating a message in the School task rather than in the Self task.

AS-unit speech rate is different from articulation rate in that AS-unit speech rate calculates the rate of speech within AS-unit boundaries alone and excludes time spent pausing in between AS-units. AS-unit speech rate was different for both tasks, but the School task (diff = -27.05) was found to have wider differences than the Self task (diff = -17.52). For the School task, the standard deviations from both collections remained almost the same (1st Collection *SD* = 60.54, 2nd collection *SD* = 58.63), while in the Self task the standard deviation decreased for the second collection (1st collection *SD* = 62.64, 2nd collection *SD* = 48.47). This indicates that in the Self task, more students were able to speak at a similar speed while the variations among learners in AS-unit speech rate for the School task did not change.

The results indicate speech rate had changed from the first to second collection for both tasks, but a significant difference was found only in the School task (+11% increase in the School task, 1% increase in the Self task).

Table 4. Descriptive statistics of speed fluency and speech quantity related measures

		First Cillection				Second Collection				
	Task	M	SD	MAX	MIN	M	SD	MAX	MIN	diff
Total response time	i	14.24	2.83	22.88	10.52	18.34	10.18	63.33	9.81	- 4.10
	s	49.75	21.91	100.41	14.39	48.56	17.63	84.15	15.14	1.20
Speech time	i	10.84	2.20	17.14	7.75	13.64	7.84	50.58	7.83	- 2.81
	s	34.03	13.73	59.58	11.90	34.69	11.43	62.40	14.03	- 0.67
Number of AS-Units	i	5.13	0.78	10.00	4.00	6.53	2.58	19.00	4.00	- 1.40
	s	9.90	3.07	17.00	5.00	10.03	3.47	17.00	3.00	- 0.14
	s	9.77	2.93	17.00	4.00	9.70	3.49	16.00	3.00	0.07
AS-Unit time	i	11.56	2.63	17.82	7.75	14.26	8.43	52.66	7.83	- 2.70
	s	40.79	17.98	78.50	12.22	40.36	14.40	69.90	14.03	0.44
Speech rate	i	268.96	43.62	351.69	146.69	273.43	42.45	373.17	186.74	- 4.47
	s	214.74	48.59	296.01	104.32	239.61	53.92	360.71	149.51	-24.87 *

*p<.05 **p<.01
i= Self introduction , s= School day

The number of AS-units indicates that for the Self task, an almost identical number of AS-units were produced in the first collection (*SD* = 0.78), since the standard deviation is less than one. Phrasal sequences were introduced in the chapter when the first set of data was collected. Thus it is likely that at the first collection point, the students did not elaborate or use original responses, but rather answered using these set phrases. At the second collection, the number of AS-units increased (*M* = 6.53, *SD* = 2.58).

For the School task, the fact that total response time decreased (- 2.4%) and speech time increased from the first to second collection (+1%) could be an indication that the students progressed in their acquisition of grammar forms and pronunciation because less time was used to articulate them. However, the number of AS-units remained the same between the collection points (diff = 0.14) indicating that the students

produced the same amount of content for both collections, which may be due to the nature of the task itself as describing a school day may tend to follow a standard format and may limit the content of the students' responses than Self task.

Summary of speed fluency and speech quantity development

In summary, these findings regarding speed and speech quantity can be interpreted as learners showing steady development in speed fluency from very early stages of their language learning. The way learners develop their speaking skills, at least when fluency is considered, in the beginning stage in language learning may be closely related to speed fluency and speech quantity.

Pause related measures

Silent pause measures. Table 5 is a summary of silent pause related measures. Silent pause ratio is the percentage of silent pauses in the total response time. The results showed a increase in silent pause time between the collections for the Self task (+41%) but not for the School task (-12 %). In the School task, silent pause ratio decreased (-9%) between the first and second data collection point. As total response time increased in the Self task (+28%), so did silent pause time. It is interesting to note changes in the location of the pauses as the semester went on. Silent pause measures within AS-units decreased (-27% in the Self task, -21% in the School task) as the course progressed, and silent pause time between AS-units increased (+12% in the Self task).

Table 5. Summary of silent pause related measures

		First Cillection				Second Collection				
	Task	M	SD	MAX	MIN	M	SD	MAX	MIN	diff
Silent pause ratio	i	22.71	6.63	34.93	4.72	24.61	8.93	48.00	9.38	- 1.91
	s	28.61	11.65	59.10	12.96	26.03	11.04	50.20	7.29	2.59
Silent pause time	i	3.25	1.12	5.80	0.50	4.61	3.06	13.78	1.52	- 1.37
	s	15.50	11.14	40.83	2.50	13.56	9.28	40.80	1.10	1.95
Mean silent pause	i	0.60	0.13	0.94	0.36	0.70	0.26	1.38	0.38	- 0.10
time	s	0.72	0.25	1.43	0.42	0.72	0.19	1.34	0.46	- 0.01
Silent pause time	i	0.67	0.82	4.15	0.00	0.57	1.03	4.31	0.00	0.11
within AS	s	6.69	6.61	25.35	0.00	5.46	6.49	29.42	0.00	1.24
Mean silent pause	i	0.33	0.23	0.83	0.00	0.24	0.01	1.08	0.00	0.09
time within AS	s	0.51	0.18	0.84	0.00	0.51	0.21	1.15	0.00	- 0.01
Silent pause time	i	2.58	0.86	4.34	0.50	4.05	2.24	10.67	1.34	- 1.47
between AS	s	8.82	5.92	21.89	2.18	8.10	4.02	20.53	1.10	0.72
Mean silent pause	i	0.66	0.16	1.08	0.39	0.74	0.32	1.83	0.38	- 0.09
time between AS	s	0.96	0.54	2.91	0.46	0.89	0.31	1.66	0.51	0.07

*p<.05 **p<.01

i= Self introduction , s= School day

What is noteworthy about this finding is that in both tasks learners seemed to pause at grammatical junctures rather than in the middle of meaningful segments of speech, which can be seen by the increases in the silent pause count between AS-units. This should be considered as a characteristic of more native-like speech. In other words, it represents development in the right direction. Increases in silent pause counts between AS-units in and of themselves, however, should not be seen as a sign of dysfluency since many of these represent a natural need to breath.

The insertion of pauses where native speakers would not normally pause is characteristic of beginning L2 learners. Silent pauses may be a good indicator of the learners' progress in their language ability. The changes in the insertion of silent pauses are worth investigating throughout all levels of L2 Japanese learners. Despite the changes in the location of pauses, the average time of pauses inserted did not change. The

duration of one semester may not be long enough to tell whether or not this will change or remain the same.

Filled pause measures. Filled pause measures did not show significant differences between the first and second collections (see Table 6). Fewer than half of the learners used filled pauses in each collection (N = 11), and only one learner in particular used filled pauses at both collection points (both collections of the School task and in the first collection of the Self task).

Table 6. Summary of filled pause related measures

		First Cillection				Second Collection				
	Task	M	SD	MAX	MIN	M	SD	MAX	MIN	diff
Filled pause count	i	0.30	0.88	4.00	0.00	0.20	0.55	2.00	0.00	0.10
	s	0.43	2.19	12.00	0.00	0.80	1.58	6.00	0.00	- 0.37
Filled pause time	i	0.11	0.32	1.54	0.00	0.08	0.23	0.90	0.00	0.03
	s	0.22	1.14	6.27	0.00	0.28	0.61	2.41	0.00	- 0.07
Filled pause count	i	0.07	0.37	2.00	0.00	0.10	0.40	2.00	0.00	- 0.04
within AS	s	0.23	1.10	6.00	0.00	0.60	1.22	4.00	0.00	- 0.37
Filled pause count	i	0.23	0.63	2.00	0.00	0.10	0.31	1.00	0.00	0.14
between AS	s	0.20	1.10	6.00	0.00	0.20	0.61	3.00	0.00	0.00

*p<.05 **p<.01
i= Self introduction , s= School day

It is therefore unlikely that this result is representative of general cases. Findings from silent pause measures indicated that the learners had begun to acquire the ability to pause at more correct positions in their sentences. The filled pause phenomena were found to reflect an increased level of complexity in their speech (Watanabe et al., 2004), as can be seen by the increase in filled pause count within AS-units for both tasks. It appears that filled pauses are not something that learners acquire but are rather a result of breakdown phenomena. That is to say, in their attempts to produce more elaborate responses, the learners devoted more attention to retrieving knowledge of different

grammar points which resulted in the increased use of filled pauses. The filled pause count between AS-units remain unchanged from the first to second collection in the School task, while it decreased in the Self task (-56%). Silent pauses within AS-units and between AS-units seemed to be a good indicator of learners' language development in terms of their ability to produce speech more efficiently. Filled pauses between AS-units occurred less often than filled pauses within AS-units (0.60 and 0.20 for the second collection). Together with silent pauses, pausing within AS-units may be indicative of their dysfluency since it disrupts the message, but pauses between AS-units may indicate different phenomena and do not necessarily indicate dysfluency.

Iwashita (2010) found in her study that lower proficiency students produced fewer pauses and repairs than higher proficiency level speakers. Likewise, the results from this study seem to indicate that the filled pause phenomenon does not occur frequently at this level. Therefore, it is important to examine this pausing phenomenon throughout a range of ability levels and to continue tracking novice level learners' progress as they reach more advanced levels.

AS-unit related measures

A summary of AS-unit related measures is shown in Table 7. These measures indicate complexity in speech. Additionally, the number of sentences divided by AS-units indicates how many AS-units were in a full sentence. As Table 7 indicates, however, the number of sentences and AS-units in this instance are almost identical (M = 4.97, M = 5.17 in Self task at 1st collection), which means learners in this study did not create sentences that contain two or more AS-units. A sentence is considered complex if it has

subordinate or coordinate clauses, as in "She went to study while I was shopping at the store." At this point in Japanese 101, no complex sentence structures had been introduced, and as a result, almost all of the sentences produced by learners were simple sentences. The differences seen in the number of sentences and AS-unit speech rate is due to some learners producing sentences that contained pauses longer than 0.5 seconds, which is a condition for separating AS-units (Foster et al., 2000).

Table 7. Summary of AS-unit measures

		First Cillection				Second Collection				
	Task	M	SD	MAX	MIN	M	SD	MAX	MIN	diff
Number of AS-Units	i	5.17	0.78	10.00	4.00	6.53	2.58	19.00	4.00	- 1.40
	s	9.90	3.07	17.00	5.00	10.03	3.47	17.00	3.00	- 0.14
Number of error-free	i	4.97	0.96	10.00	3.00	6.33	2.70	19.00	4.00	- 1.37
AS-Units	s	8.47	2.69	13.00	4.00	8.93	3.33	16.00	1.00	- 0.47
Number of sentences	i	5.17	0.75	7.00	4.00	6.50	2.60	19.00	4.00	- 1.34
	s	9.77	2.93	17.00	4.00	9.70	3.49	16.00	3.00	0.07
AS-Unit time	i	11.56	2.63	17.82	7.75	14.26	8.43	52.66	7.83	- 2.70
	s	40.79	17.98	78.50	12.22	40.36	14.40	69.90	14.03	0.44
AS-Unit speech rate	i	333.78	62.64	487.77	157.20	351.29	48.47	467.15	244.60	-17.52
	s	257.74	60.54	375.30	139.81	284.79	58.63	389.09	176.99	-27.05

*p<.05 **p<.01
i= Self introduction , s= School day

Number of AS-units (diff = -1.40) and AS-unit time (diff = -2.70) increased in the Self task. But in the School task, these measures did not show as much change as in the Self task. AS-unit measures by themselves have limitations in that it doesn't take into account what is said. However, from the fact that silent pauses decreased and filled pauses were observed from only a few learners, it can be surmised that the learners were able to make longer sentences instead of taking more silent pauses when articulating an AS-unit. By observing the means and standard deviations in both tasks, it seems that in the School task the participants responded using various numbers of AS-units starting

from the first data collection point, but no increases were seen in the second collection. In other words, each participant answered differently for the School task, whereas in the Self task, they answered more uniformly, perhaps owing to the more formulaic nature of the discourse.

Error-free AS-units. Error-free AS-units improved between the first collection and second collection for both tasks (see Table 7), but the differences were not significant for the School task. The percentage of error-free AS-units was calculated by dividing the number of error-free AS-units by the total number of AS-units. The results showed that the percentage remained at 97% for both collections of the Self task. For the School task, it increased slightly from 86% to 89%. The number of error-free AS-units alone cannot indicate accuracy of speech, but when combined with the total number of AS-units, a subtle change was seen.

Repair Fluency

Repair fluency measures did not produce any significant findings in the tasks for either collection point. The only exception was that self-correction count in the School task decreased in the second collection, but due to the small number of learners (N = 8 in the 1st collection, N = 5 in the 2nd collection) who made self-corrections, this finding cannot be considered generalizable.

Table 8. Summary of Means and Standard Deviations

		First Cillection				Second Collection				
	Task	M	SD	MAX	MIN	M	SD	MAX	MIN	diff
Repeat count	i	0.07	0.25	1.00	0.00	0.03	0.18	1.00	0.00	0.04
	s	0.10	0.40	2.00	0.00	0.07	0.25	1.00	0.00	0.04
Stutter count	i	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	s	0.13	0.57	3.00	0.00	0.17	0.53	2.00	0.00	- 0.04
Self-correction count	i	0.03	0.18	1.00	0.00	0.03	0.18	1.00	0.00	0.00
	s	0.80	1.67	7.00	0.00	0.23	0.50	2.00	0.00	0.57

*p<.05 **p<.01
i= Self introduction , s= School day

Of all the learners who made self-corrections in both collections, only two were the same learners. Consequently, it cannot be said that the repairs are due to individual characteristics. Stuttering was not found at all in the self-introduction task, which could be due to the characteristics of the task itself, as the phrases used were very formulaic and more easily acquired than the phrases learners used to complete the School task.

Task differences and speech elicitation

The two tasks used in this study were chosen for two main reasons. The first is that both tasks were introduced at an early stage in the Japanese 101 language course. This is crucial when novice level learners are targeted for a fluency study because fluency cannot be measured if the tasks were too difficult. The second reason is that the tasks can be reused at a more advanced level with potential for learners to use more elaborative speech as their ability increases.

For this study, these two tasks showed different results with regard to the quantity and speed of speech. For quantity, the Self task showed larger gains, but when speed is considered, the School task improved slightly more than the Self task. When looking at changes in mean and the differences in standard deviation between the tasks, it is clear

that task characteristics influence students' motivation to speak using more sentences. In the Self task, the self-introduction, learners seemed to make the same number of sentences to complete the task since the standard deviation was less than 1 at first collection point (See Table 7). In the second collection, however, as the mean of the number of AS-units went up, there seemed to be greater variation among the learners. The table backs this up with larger standard deviation values.

On the other hand, leaners did not seem to engage in longer responses in the School task as the mean of number of AS-units did not increase as much as in the Self task. The standard deviation also remained nearly the same. This shows that the School task did not encourage students to produce more sentences than the self-introduction task did despite the larger variety of sentences and response time as compared to the first collection point.

The School task did not seem to encourage students to make more original speech since the number of AS-units remained the same while silent pause ratio decreased. The decrease in silent pause could mean that the learners were able to make sentences more efficiently. These findings tend to show that learners were encouraged to say a greater variety of things in the self-introduction task than in the school-days task. It is interesting to see what kind of tasks would elicit more speech and motivate students to use more elaborate and original sentences.

Summary

Of all the objective measures analyzed in this study, speed related measures and speech quantity related measures showed gains between the first and second collection

points for both tasks, although significant differences were only found in Speech rate and Mean run length. Pause related measures did not show any significant differences between the collection points, but the learners tended to pause at more native-like locations (i.e. between AS-units) at the second collection point in both tasks. The mean pausing time did not show development in either task, which could mean that for novice learners the location of pauses develops more quickly, but the length of pause takes longer to improve.

Research question 2

In this section objective measures are analyzed in order to answer the second research question: Which objective measures are correlated with subjective ratings? The School day task was used to examine correlation with subjective ratings, and both the first and second collection were analyzed to determine how changes in objective measurements correlated with subjective rating. Table 9 shows descriptive statistics of the subjective ratings obtained by Japanese instructor for both collection points. Inter-rater reliability among the three Japanese instructors as measured by average pairwise percent agreement was 43.33%.

Table 9. Descriptive statistics of subjective rating

Subjective rating	1st collection Mean (SD)	2nd collection Mean (SD)
	88.80 (5.43)	90.72 (6.22)

Table 10. Correlation coefficients between speed fluency measures and subjective ratings

	1st Collection	2nd Collection
Speech rate	0.60*	0.65*
AS-Unit speech rate	0.63*	0.70*
Articulation rate	0.41	0.57**
Mean run length	0.54*	0.50*

*p<.05, **p<.01

Speed fluency, quantity of speech, and subjective ratings

Speech rate (r = 0.60, r = 0.65) and AS-unit speech rate (r = 0.63, r = 0.70) showed stronger correlations than the other measures shown in Table 10. Both measures showed stronger correlations at the second collection point. Articulation rate, which excludes silent pauses, correlated relatively weakly at the first collection point (r = 0.41), but it correlated more strongly at the second collection point (r = 0.57). These three measures are all speed fluency measures and they all correlated more strongly at the second collection point.

Mean run length correlated to subjective ratings to nearly the same extent at both the first (r = 0.54) and second collection (r = 0.50). Mean run length, unlike other speed fluency measures, was not affected by students' level progression and can be said to have correlated steadily to subjective ratings. Of these measures, the AS-unit speech rate correlation remained the strongest for both the first and second collection point (r = 0.63, r = 0.70).

Table 11. Correlation between subjective ratings and speech quantity measures

	1st Collection	2nd Collection
Number of AS-Units	-0.01	0.25
Speech time	-0.13	0.27
Total number of syllables	0.06	0.42

*p<.05, **p<.01

Of all the speech quantity measures, the number of AS-units (r = -0.01, r = 0.25) and speech time (r = -0.13, r = 0.27) correlated very weakly at the first collection but it correlated slightly more strongly at the second collection point. The correlation for the total number of syllables was much higher at the second collection point than at the first (r = 0.06, r = 0.42).

What is interesting in these findings is that rate of speech , rather than speech quantity, appears to be an integral part of judging proficiency for this level of learners, if not the most important part.

Silent pause measures and subjective ratings

Table 12. Correlation between subjective ratings and silent pause measures

	1st Collection	2nd Collection
Silent pause ratio	-0.52*	-0.50*
Average silent pause time within AS	-0.54*	-0.48*
Average silent pause time	-0.39	-0.37

*p<.05, **p<.01

Silent Pause Ratio and Average Silent Pause Time. Average silent pause time (r = -0.39, r = -0.37) did not have as strong a correlation as silent pause ratio (r = -0.52, r = -0.50). This could mean that at this level (Japanese 101), the length of each pause is not

so much of a concern, but the frequency of pauses, as opposed to overall time to answer the question, was crucial. This can change as participants move into higher levels, as well. However, when only pausing within AS-units is considered, it is correlated strongly to subjective ratings. This may be due to the fact that pauses are major distractions for raters when listening to speech. Average silent pause time within AS-units correlated about the same for both collection (r = -0.54, r = -0.48). When compared with average silent pause time that included all of the silent pauses within the audio, average silent pause time within AS-units had a stronger correlation value.

Table 13. Correlation between subjective ratings and silent pause measures

	1st Collection	2nd Collection
Silent pause time (Total)	-0.58*	-0.34
Silent pause time within AS	-0.68*	-0.45
Silent pause time between AS	-0.33	-0.06

*p<.05, **p<.01

All of these measures in Table 13 correlated very differently at the second collection point as oppose to the first, and they decreased in all cases. Silent pause time (total) correlated more strongly at the first collection point (r = -0.58) than at the second (r = -0.34). Silent pause time within AS-units correlated the most strongly to subjective ratings for the first collection (r = -0.68) but declined to r = 0.45 for the second collection.

When comparing the first and second collection points, silent pause related measures that calculated the total pause length or count seemed to have a weaker correlation at the second collection point. Silent pause ratio and average silent pause time within AS-units were the only measures that maintained a relatively strong correlation to

subjective scores. For the first collection, the longer the aggregate pause time and the more frequent the pausing, the lower the students' proficiency was rated. For the second collection, as the students learned to pause at correct locations, pausing no longer seemed to correlate to proficiency ratings. Consequently, pausing could be a characteristic that can help distinguish learners in the very beginning stage of language acquisition from those at later stages.

Silent pause time within and between AS-unit. Silent pause time within AS-units (r = -68, r = -45) showed stronger correlations than Silent pause time between AS-units (r = -0.33, r = 0.06) at both the first and second collection points. This may indicate that when the raters judged proficiency, pauses that occurred between AS-units were not recognized as dysfluency or as something that negatively impacts proficiency. This suggests that as long as pauses are located at grammatical junctures, they may not affect proficiency ratings.

Silent pause time (total) and silent pause within AS-unit. Between the silent pause time (total) (r = -0.58, r = -0.34) and silent pause time within AS-units (r = -0.68, r = 0.45), the silent pause time within AS-units correlated more strongly with subjective ratings. However, the correlation of both measures of silent pauses declined at the second collection point. This could be the result of learners pausing at more native-like locations, causing the raters to not pay attention to pauses within AS-units.

Changes in pausing and subjective rating. There have been studies targeting higher level learners that found silent pause ratio and average pause time to be indicators of proficiency (Iwashita et al., 2008; Lennon, 1990; Taguchi & Iwasaki, 2008). When compared to speed fluency measures, the correlation with pause related measures

declined at the second collection point in this study. The findings from Research Question 1 support the idea that the learners have developed the ability to pause at more correct locations when comparing silent pauses within AS-units to silent pauses between AS-units. It is therefore reasonable to say that learners' acquisition of pausing at correct locations corroborated the findings that pausing had less of an influence at the second collection point. Not only that, speed related measures continued to have a stronger correlation at the second collection point. This could mean that the raters paid more attention to speed at both collection points regardless of the learner's level.

AS-unit error-free AS-unit ratio. The error-free AS-unit ratio (r = 0.59 for the first collection) correlated more strongly at the second collection point (r = 0.77). Judging a speaking performance is very complex as the raters must pay attention to many different aspects of the performance. In this study, a standardized proficiency rating scale was created in order to obtain consistency among the raters. However, the correlation values indicate that some aspects of speaking may affect the subjective ratings to varying degrees depending on the time data was collected. The aspect that showed the greatest variation was the error-free AS-unit number.

Table 14. Correlation between subjective ratings and Error-free AS-unit ratio

	1st Collection	2nd Collection
Error-free AS-unit ratio	0.59*	0.77*

*p<.05, **p<.01

Overall, speed fluency retained a stronger correlation to subjective ratings at both the beginning and end of the semester. However, pause related measures correlated more weakly at the second collection point. As the learners' levels increased, the relationship

between proficiency and the objective measures can change, but this could not be seen in a semester-long study. As the correlation between subjective ratings and the objective measures showed different results at the first and second collection points, different correlations may arise when examining more various proficiency levels.

Filled pauses. There were not many filled pauses found in the first collection, which could be the reason for the weak correlations overall. In the second collection, however, filled pause related measures within AS-units correlated more strongly than at the first collection.

Table 15. Filled pause measures and correlations

	1st Collection	2nd Collection
Average filled pause time	-0.08	-0.35
Filled pause time within AS	-0.10	-0.37
Filled pause time between AS	-0.10	-0.00
Average filled pause time within AS	-0.08	-0.36

*p<.05, **p<.01

Although a total of 9 learners used filled pauses, only 2 participants used filled pauses during the first collection. The fact that fewer than half of the participants in this study used filled pauses means that this correlation cannot be said to be representative of learners in general. Again, this could be different for different language levels as more participants in this study use filled pauses in the second collection. The use of filled pauses could increase in tandem with language level since the cognitive load of producing more complicated sentences could manifest itself as more frequent uses of filled pauses as learners buy time to organize their output. It is worth investigating further with learners at a variety of levels in order to analyze how filled pauses influence the raters' judgment of proficiency.

Proficiency score range and objective measures

In this section, objective measures are analyzed based on the proficiency score range of the students. The objective measures used in this section are those which showed relatively strong correlations to the subjective ratings. They are AS-unit speech rate, mean run length, and pause ratio.

AS-unit speech rate and proficiency. AS-unit speech rate (ASSR) was used to analyze the characteristics of those who have higher proficiency ratings since this measure correlated the most to the subjective ratings of all the objective measures. As seen in Figure 1, the subjective score range and ASSR range were nearly parallel.

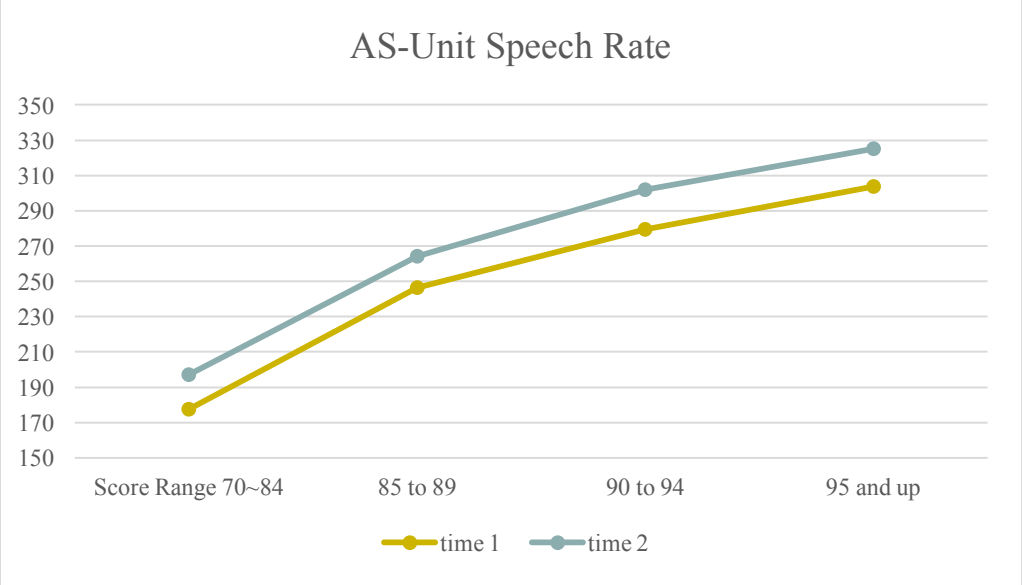

Figure 4. AS-unit speech rate and the score range

The findings were that those with an ASSR in the high 100s received lower subjective ratings (70 ~ 84) and those who received higher subjective scoring (90 and above) had an ASSR in the low to mid 300s. It is also clear that from the first collection

to second collection, learners were able to speak faster. The score ranges and differences in ASSR lining up clearly could mean that even though the correlation value was between .60 and .70, this could be used as a good indicator to use to estimate proficiency. As the correlation coefficients reveal and many past studies have concluded, there is not any single measure good enough to determine proficiency, and the measures need to be analyzed in relationship to other measures due to the nature of speech production.

Mean run length and proficiency. Figure 2 shows the correlation between mean run length (MRL) and the subjective score ranges.

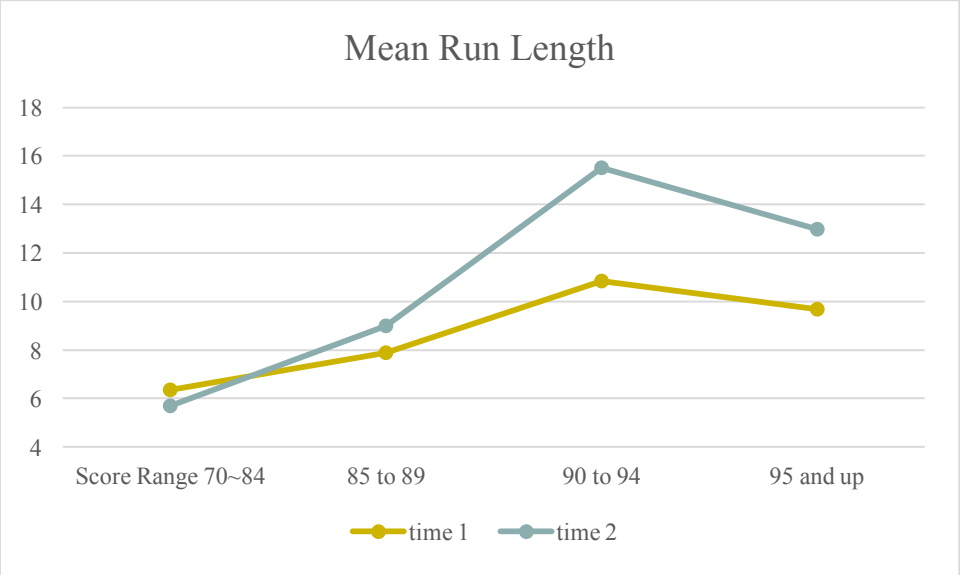

Figure 5. Mean run length and the score range

As the correlation values suggest, mean run length (r = .50 to r = .54) is not as neatly parallel with the subjective ratings as was ASSR. What can be said is that the students are clearly divided by MRL except those in the score range above 95. Those who had MRL of less than 10 stayed in the lower to mid scoring range (70~89), and those with MRL of 10 or more, usually closer to 20, were in the higher scoring range (90 and

up). This can mean that the MRL does not necessarily reflect one's proficiency level due to the lack of the relationship with accuracy that was found to be a major contributor to receiving a higher score in this study. The run could be long, but if it contains more errors and pauses it may not result in higher scores.

Silent pause ratio and proficiency. Although the correlations were not particularly strong (r = -0.52, r = -0.50), higher proficiency students (90 and up) tended to have a steady pause ratio in the 25 to 30 range.

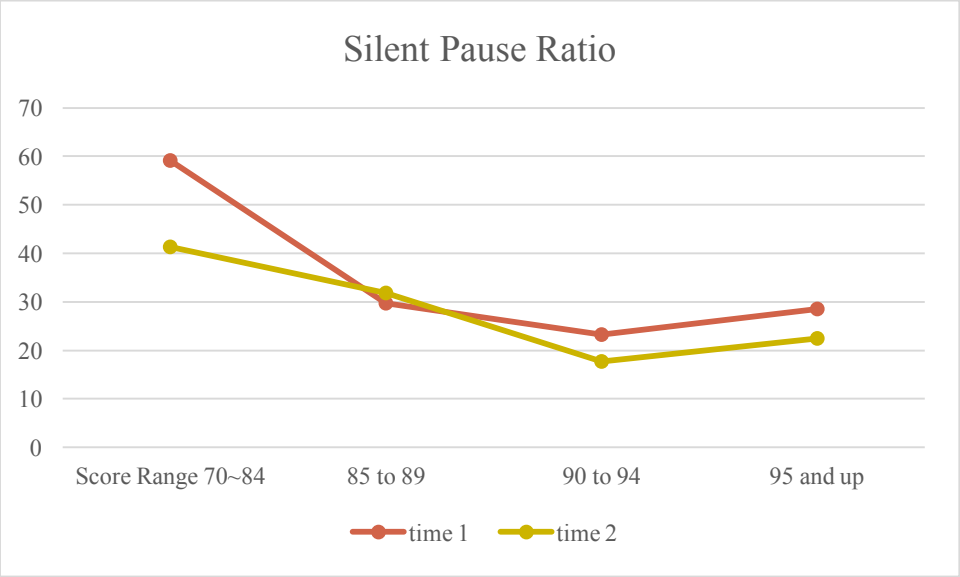

Figure 6. Silent pause ratio and score range

It is interesting to note that a score of less than 25 did not seem to correlate with the higher subjective ratings, and comparatively higher score students had a pause ratio of between 25 and 30. This could mean that speaking with fewer pauses does not lead to better subjective ratings, as it may not be required in order to deliver a message efficiently or be easily understood. According to the results presented here, a pause rate of about one fourth of the total speech length may be optimal.

CHAPTER 5. CONCLUSION

Introduction

This study investigated how the speech of novice L2 Japanese learners in a classroom setting would change over the course of a semester in terms of objective measures and how these objective measures and subjective ratings of their speech obtained from instructors correlated to each other. This final chapter will summarize the research findings and discuss their pedagogical implications, limitations of the study, and directions for future studies.

Research findings

Development of speech production over a semester. Statistical analysis indicated that speech rate changed significantly from the first collection to the second collection. AS-unit speech rate changed the most out of all the measures although the change was not significant. Articulation rate also changed but not as much as AS-unit speech rate. The difference between speech rate and articulation rate is the inclusion of silent pauses, and consequently, the results are closely related to pausing patterns in learners' speech.

Speech quantity measures, total response time and speech time, showed gains in the Self task, which means that learners were able to produce large utterances at the

second collection point. In the School task, learners did not seem to make longer responses at the second collection, but they spent less time responding. This can also be seen from the fact that total response time decreased while speech time increased. A trend was seen in several students where less time was spent to produce almost the same set of sentences, and this indicates progress in their fluency. The number of AS-units indicated that there was no decrease from the first collection to the second collection, so it is reasonable to come to the conclusion that they were able to speak faster while responded with the same number of sentences at the second collection point.

These findings regarding speed and speech quantity can be interpreted as learners showing steady development in speed fluency from very early stages of their language learning. The way learners develop their speaking skills in the beginning stage of language learning, at least when fluency is considered, may be closely related to speed fluency and speech quantity.

Pausing measures showed that learners became able to pause at more grammatically correct junctures as the semester moved on. This can be seen by an increase in silent pause time between AS-units and a decrease in silent pause within AS-units at the second collection point for both tasks. The time spent on pausing shifted from within the AS-units to between the AS-units at the second collection point. It is interesting to note that as the semester went on, the use of filled pauses increased. This can be seen as an increase in the level of complexity in their speech as filled pauses are generally an indication of struggling to form a sentence. However, this result may not be applicable to general cases due to the small number of learners who used filled pauses.

Tasks and objective measures. This research used two tasks to collect audio data. The tasks were chosen for their flexibility in elaboration according to ability level while still being applicable to novice level learners. As a result, the objective measures obtained from these two tasks revealed very different measurements, though the measurements from these two tasks were affected by the characteristics of the tasks and their relationship to the students' motivation for speaking. The first finding is that the Self task elicited more various responses from the students than the School task did. This is seen in increases in speech time, the number of AS-units, and the total number of syllables. The standard deviations from the Self task reveal that at the second collection point, that there was a wide range in the total number of sentences produced by the learners. While the Self task encouraged students to speak more at the second collection point than at the first, it did not elicit faster speech than the first collection. The second finding was that for the School task, while no significant change was seen in terms of the number of AS-units, speech rate became faster at the second collection point which was significantly different between the collections, but the difference was only subtle in the Self task.

The flexibility of the Self task revealed itself in how the learners responded, and as a result, quite different measurements were found between the two tasks. These tasks should be further examined in order to find the best task to elicit speech for fluency research.

Correlation analysis. Correlation coefficients between subjective ratings and objective measures were calculated to see which measures correlate strongly with subjective ratings given by Japanese instructors. The results were that speed fluency

measures correlated strongly at both the first and the second collection points, and the correlation grew stronger at the second collection. Speech quantity measures such as the total number of moras did not show as much correlation as speed fluency measures, but at the second collection point, the correlation was found to be stronger than at the first collection. At the beginning of the semester, none of the learners responded with a substantial number of sentences, and that may have been reflected in the correlation value.

Silent pause measures showed weaker correlations at the second collection point, especially those concerned with the location of pauses. It can be said that silent pause ratio could also be an integral part of fluency since it retained roughly the same strong correlation at both collection points. An increase in silent pauses between AS-units and a decrease in silent pauses within AS-units may be the reason the correlations weakened at the second collection point. It may be that as learners spent less time pausing at unnatural locations, the raters' attention focused on other parts of speech production such as accuracy and complexity features. In fact, when error free AS-units was also analyzed, the measure correlated strongly at the second collection point even more so than speech rate.

There were not enough filled pauses to warrant generalizations, but the increase in filled pauses as the semester went on may have affected correlation values. Although the correlation was weaker at about r = -0.37 for filled pause time within AS-units, analysis found that occurrences of filled pauses negatively correlated to subjective scoring. It is possible that filled pauses may continue to be negatively correlated to subjective scores

as learners increase in ability. This should be further examined in a longer-term study in the future.

Although data was collected only twice in a semester, this study revealed how the way raters judged speaking performances changed by means of changes in correlation values. At the very beginning level, ratings might have been related to the location of silent pauses, but as learners' skills increased, the length of overall speech also began to relate to the ratings. Speech rate, pause ratio, and mean length of run remained important for subjective ratings regardless of the learners' level.

This study did not uncover whether or not repair affects subjective ratings because repairs were minimal in the data collected. Repair can be a characteristic of spontaneous speech, and monologic tasks such as the ones employed in this study do not generate repairs due to the nature of the tasks. A different type of task may need to be used in order to analyze the effect of repair.

Pedagogical implications

The developments seen in the research findings suggest that the ability to speak at a certain speed, with a certain length between pauses, and with pausing at grammatical junctures is key to receiving better ratings. Consequently, fluency can be developed by pushing production toward longer segments when practicing as this can encourage students to pause less often within meaningful segments. The fact that speech rate and other speed related measures remained an integral part of the subjective ratings throughout the semester means instructing learners to speak with a certain level of speed

and to maintain that speed may be beneficial from very early on in the language learning process.

Individual differences also need to be taken into account, but it is crucial to include not only grammatical instruction but the ability to speak at a certain rate in classroom instruction in order to avoid certain dysfluent speaking styles from becoming engrained.

Learners must also be instructed that simply talking very fast does not lead to fluency, but rather pausing for the right amount of time at appropriate locations is a key to enhancing fluency. The number of filled pauses increased as the learners increased in ability, and filled pauses within AS-units, in particular, were negatively correlated to the subjective ratings, although the correlation was weak at the first collection point. The number of filled pauses may continue to increase later in the language learning process, at which point in-class instruction on how to use filled pauses and how to avoid using them extensively may become necessary, as this may interrupt meaningful segments of speech.

When learners formulate speech at the beginner level, they need to pay attention to not only grammar forms but also pronunciation and the content of the message. These components of speech can largely affect how efficiently they can speak, and thus impact fluency. The fewer components learners need to pay attention to, the more they can focus on delivering a message in a fluent manner, which is represented by speed related measures. Therefore, the repeated practice of forms and pronunciation can be beneficial when teaching fluency so that learners can produce speech with less cognitive load placed on these fundamental parts of speech.

Limitations of the present study

Length of observation. This study examined speech production over the course of one semester. The results indicated that the speech production of novice learners changed from the beginning to the end of the semester. However, in order to analyze and identify how fluency develops across all levels of learners, observations that span across multiple semesters are required. It becomes difficult to track the same students for more than a semester as the number of students in Japanese courses generally decreases as the level of difficulty increases, but this will be a necessary step in order to establish fluency measurements for each level of ability.

Complexity. In addition, the learners examined in this study were all novice level, and at this point complexity factors were absent from their speech production. Therefore, it is beneficial to analyze speech samples from advanced level learners to understand how fluency can be seen at higher levels of complexity.

Task Design. There were two tasks used in this study, and the measures showed different results between the tasks. It could be that the characteristics of each task were quite different even though both had flexibility and both came from early chapters in the course. It is not clear which task elicited fluency characteristics representative of their overall language use.

Future directions

Various proficiency levels. More data from various levels of proficiency in Japanese L2 learners is needed in order to see how these objective measures behave at the advanced level. These measures could show different developmental stages in the long

process of language acquisition. Longitudinal observation of a learner from the beginning of their language learning to the advanced level may provide valuable information as to how fluency develops. Establishing which measures can be used as proficiency guides can be accomplished not only by tracking the same student but also by analyzing speech production across all ability levels.

Fluency focused instruction. It is clear that some aspects of fluency (speech rate, articulation rate, AS-unit speech rate) interacted with subjective scoring, and instructing students to improve fluency will contribute to the achievement of proficiency. However, methods of fluency instruction and the effects of this instruction has yet to be seen. Measuring the development of learners' fluency by using the objective measures that correlated with subjective scores would provide the knowledge of what exactly can enhance fluency in these classroom activities. The activities that contributed to fluency development for the novice learners in this study can serve as a guide to creating fluency focused instruction.

Fluency differences between learners' L1 and L2. There were clear trends seen in the correlation between the subjective and objective scores, but there may also be differences in the native languages of the learners, as well as individual speaking styles that could have positively or negatively influenced these scores apart from linguistic ability. Identifying the inherited characteristics in their L1 and those of Japanese native speakers can further aid instructors on how to instruct students to speak more similarly to Japanese native speakers. Just as pronunciation can be largely affected by their first language, pausing and speaking speed could be largely derived from their L1, as well.

Identifying these characteristics may be the first step in understanding how best to instruct fluency as there could be some that negatively impact subjective ratings.

Final remarks

The results of this study suggest that an increase in speech quantity is a major contribution to the changes seen in novice level learners. Changes in the location of pauses were also seen as the semester went on, and pause placement became more similar to how native speakers would pause. Filled pauses were not seen very much at this level of learners, but the results suggest that the number of filled pause in the learner could increase in parallel to their level. Many previous fluency studies have focused on advanced learners of ESL or EFL, but this study suggests that fluency development is also seen in novice level learners, and fluency is not a characteristic only seen in advanced learners. The way in which advanced learners achieve their fluency and what kind of obstacles they may encounter along the way can be captured with further longitudinal studies of fluency. The possibility of predicting proficiency using objective measures has also been studied while targeting mostly advanced learners, but this study has revealed that speech samples from novice learners indicate similar results as those obtained with advanced students, which is that speech rate can be a good indicator of proficiency. As many studies have suggested, no single measure can be extrapolated to capture the whole dynamic of speaking performance, and therefore it is necessary to investigate learners at a wide variety of ability levels in order to determine which measures correlate to subjective ratings across all levels. The correlation coefficient values revealed that the speed of speech does play an integral part in subjective ratings

regardless of level, and instructing fluent speech from early on can help learners gain better proficiency scores.

LIST OF REFERENCES

LIST OF REFERENCES

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.

Boersma, Paul & Weenink, David (2016). Praat: doing phonetics by computer [Computer program]. Version 6.0.16, retrieved 5 April 2016 from http://www.praat.org/

Chambers, F. (1997). What Do We Mean By Fluency?, *25*(4), 535–544.

Council of Europe. (2001). the Common European Framework of Reference for Languages: Learning, Teaching, Assessment. *Council of Europe*, 1–273.

De Jong, N. H., Groenhout, R., Schoonen, R., & Hulstijn, J. (2013). Second language fluency : Speaking style or proficiency ? Correcting measures of second language fluency for first language behavior. *Applied Psycholinguistics*, *36*(2), 1–21.

De Jong, N. H., & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior research methods*, *41*(2), 385-390.

Fillmore , C. J . (1979). On fluency. In C. J. Fillmore, D. Kemple , & W. Wang (Eds.), *Individual differences in language ability and language behavior* (pp. 85 – 101). San Diego, CA: Academic Press .

Foster, P., & Skehan, P. (1996). The Influence of Planning and Task Type on Second Language Performance. *Studies in Second Language Acquisition*, *18*, 299.

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics*, *21*(3), 354-375.

Freed, B., & Ebrary, Inc. (1995). *Second language acquisition in a study abroad context* (Studies in bilingualism ; vol. 9). Amsterdam ; Philadelphia: J. Benjamins.

Fujimori, C., & Koizumi, R. (2011). Correlations Between Holistic Ratings of Speech Presentation by the ALT and the JTE, and Objective Measures and Students' Peer Evaluation. *KATE Journal*, (25), 21–31.

Fukada, A. (2013). An online oral practice/assessment platform: Speak Everywhere. *The IALLT Journal*, 43(1), 64-77

Fukada, A., Hirotani, & M., Matsumoto, K. (2015) Fluency Calculator (Version 1.0.2) A Tool for Calculating Fluency–related Measures [Software].

Ginther, A., Dimova, S., & Yang, R. (2010). Conceptual and empirical relationships between temporal measures of fluency and oral English proficiency with implications for automated scoring. *Language Testing*, *27*(3), 379–399.

Ishizaki, A. (2004). What Effect Do Pauses Have on Listenability?: A Reaction to Recitations by a Native Japanese Speaker and by a Learner of Japanese as a Second Language. *Gengo Bunka to Nihongo Kyouiku. 2004*, 90–101.

Ishizaki, A. (2005). How Does a Learner Leave a Pause When Reading Japanese Aloud?: A Comparison of English, French, Chinese and Korean Learners of Japanese and Native Japanese Speakers. *Japanese-Language Education around the Globe ; Japanese Language Education around the Globe*, *15*, 75–89.

Iwashita, N. (2006). Syntactic Complexity Measures and Their Relation to Oral Proficiency in Japanese as a Foreign Language. *Language Assessment Quarterly: An International Journal*, *3*(2), 151–169.

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.

Iwashita, N. (2010). Features of oral proficiency in task performance by EFL and JFL learners. *Selected Proceedings of the 2008 Second Language Research Forum*, 32–47.

Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System*, *32*(2), 145–164.

Larsen-Freeman, D. (2006). The emergence of complexity, fluency, and accuracy in the oral and written production of five Chinese learners of english. *Applied Linguistics*, *27*(4), 590–619.

Lennon, P. (1990). Investigating Fluency in EFL: A Quantitative Approach. *Language Learning*, *40*(3), 387–417.

Miyamoto, M (2016). Flowchart Scoring System. Unpublished manuscript, School of Languages and Cultures, Purdue University, West Lafayette, USA.

Negishi, J. (2012). Relationships between L2 Speakers' Development and Raters' Perception on Fluency in Group Oral Interaction, *15*(2), 1–26.

O'Connell, D. C., & Kowal, S. (2008). Communicating with One Another, 175–186.

Riggenbach, H. (1991). Toward an understanding of fluency: A microanalysis of nonnative speaker conversations. *Discourse Processes*, *14*(4), 423–441.

Segalowitz, N. (2010). *Cognitive Bases of Second Language Fluency*. Hoboken: Taylor and Francis.

Skehan, P., & Foster, P. (1999). The Influence of Task Structure and Processing Conditions on Narrative Retellings. *Language Learning*, *49*(1), 93–120.

Taguchi, N., & Iwasaki, Y. (2008). Training Effects on the Development of Oral Fluency in L2 Japanese. *Japanese Language and Literature*, *42*(2), 413–437.

Takiguchi, H. (2004). A Study of the Developments of Speaking Ability Among Japanese Junior High School Students from the Viewpoints of Fluency, Complexity, and Accuracy. *KATE Bulletin*, (18), 1–13.

Towell, R., Hawkins, R., & Bazergui, N. (1996). The Development of Fluency in Advanced Learners of French. *Applied Linguistics*, *17*(1), 84–119.

Watanabe, M., Den, Y., Hirose, K., & Minematsu, N. (2004). Clause types and filled pauses in Japanese spontaneous monologues. *Proceedings of the 8th International Conference on Spoken Language Processing*, 905–908.

Zellner, B. (1994). Pauses and the temporal structure of speech. *Zellner, B.(1994). Pauses and the temporal structure of speech, in E. Keller (Ed.) Fundamentals of speech synthesis and speech recognition.(pp. 41–62). Chichester: John Wiley.*, 41–62.

Uchida, T. (2005). Impression of Speaker's Personality and the Naturalistic Qualities of Speech : Speech Rate and Pause Duration. *The Japanese Journal of Educational Psychology*, *53*(1), 1–13.

APPENDIX

APPENDIX

CHECK		
	+2	Well organized, exemplary response to the task
	+1	Shows confidence/emotion in speech
	+2	No mistake on newly learned grammar/vocab
	+1	Natural intonation, pronunciation (maybe with some L1 characteristics but consistently intelligible)
	+2	The speech is directed towards listeners (not reading a script but owning the speech)
	+1	Can produce long and complex sentences
	+1	No mistake on previously learned grammar/vocab

CHECK		
	+1	Can demonstrate some newly learned grammar
	+1	Can demonstrate some newly learned vocab
	+2	Overall, the speech is intelligible
	+2	Stay on topic throughout the speech, clear and straight to the point
	+1	Can recognize own mistakes
	+1	Can make successful self-corrections
	+1	Can demonstrate memorized phrases with only minor mistakes
	+2	Makes an effort to elaborate the speech
	+1	The speech is fluid.
	+2	Listening does not require much effort
	+1	Can demonstrate some appropriate intonation/pitch

CHECK		
	+2	Makes an effort to complete the task
	+2	Can produce (at least) 1 or 2 meaningful sentences
	+2	Can correctly use previously learned grammar
	+2	Can correctly use previously learned vocab
	+2	Stay on topic throughout the speech
	+1	Speak with confidence
	+1	Can speak with some speed.
	+2	Can talk with constant speed (may be slow but at least constant)