CAPTURING L2 ORAL PROFICIENCY WITH CAF MEASURES AS PREDICTORS OF THE ACTFL OPI RATING

by

Mayu Miyamoto

A Dissertation

Submitted to the Faculty of Purdue University
In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

School of Languages and Cultures
West Lafayette, Indiana
May 2019

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Atsushi Fukada, Chair
School of Languages and Cultures
Dr. Mariko Wei
School of Languages and Cultures
Dr. Jessica Sturm
School of Languages and Cultures
Dr. April Ginther
Department of English

Approved by:
Dr. Jennifer William
Head of the Graduate Program

ACKNOWLEDGMENTS

First and foremost, I would like to express my sincere gratitude to my advisor, Professor Atsushi Fukada, for his continued support and encouragement throughout my Ph.D. journey. Without his guidance, I could not have completed this dissertation. I could not have imagined having a better advisor for my Ph.D. study.

Besides my advisor, I would like to thank all of my outstanding mentors on my committee: Dr. Mariko Wei, Dr. April Ginther, and Dr. Jessica Sturm, for their insightful comments and encouragement for future research.

My gratitude is extended to the fellow graduate students in the School of Languages and Cultures (SLC) and in the Second Language Studies (SLS) program, who have inspired and motivated me to keep going. I am grateful to have met, worked with, and learned from them.

Last but not least, I would like to express special thanks to Mike Chen and to my family for their unconditional support, encouragement, and advice throughout the completion of my graduate program and my life in general.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Author: Miyamoto, Mayu. PhD
Institution: Purdue University
Degree Received: May 2019
Title: Capturing L2 Oral Proficiency with CAF Measures as Predictors of the ACTFL OPI Rating
Committee Chair: Atsushi Fukada

Despite an emphasis on oral communication in most foreign language classrooms, the resource-intensive nature (i.e. time and manpower) of speaking tests hinder regular oral assessments. A possible solution is the development of a (semi-) automated scoring system. When it is used in conjunction with human raters, the consistency of computers can complement human raters' comprehensive judgments and increase efficiency in scoring (e.g., Enright & Quinlan, 2010). In search of objective and quantifiable variables that are strongly correlated with overall oral proficiency, a number of studies have reported that some utterance fluency variables (e.g., speech rate and mean length of run) might be strong predictors for L2 learners' speaking ability (e.g., Ginther et al., 2010; Hirotani et al., 2017). However, these findings are difficult to generalize due to small sample sizes, narrow ranges of proficiency levels, and/or a lack of data from languages other than English. The current study analyzed spontaneous speech samples collected from 170 Japanese learners at a wide range of proficiency levels determined by a well-established speaking test, the American Council on the Teaching of Foreign Languages' (ACTFL) Oral Proficiency Interview (OPI). Prior to analysis, 48 *Complexity, Accuracy, Fluency* (CAF) measures (with a focus on fluency variables) were calculated from the speech samples. First, the study examined the relationships among the CAF measures and learner oral proficiency assessed by the ACTFL OPI. Then, using an empirically-based approach, a feasibility of using a composite measure to predict L2 oral proficiency was investigated. The results revealed that *Speech Speed* and

Complexity variables demonstrated strong correlation to the OPI levels, and moderately strong correlations were found for the variables in the following categories: *Speech Quantity, Pause*, *Pause Location* (i.e., Silent pause ratio within AS-unit), *Dysfluency* (i.e., Repeat ratio), and *Accuracy.* Then, a series of multiple regression analyses revealed that a combination of five CAF measures (i.e., Effective articulation rate, Silent pause ratio, Repeat ratio, Syntactic complexity, and Error-free AS-unit ratio) can predict 72.3% of the variance of the OPI levels. This regression model includes variables that correspond to Skehan's (2009) proposed three categories of fluency (speed, breakdown, and repair) and variables that represent CAF, supporting the literature (e.g., Larsen-Freeman, 1978, Skehan, 1996).

Keywords: oral proficiency, assessment, automated scoring system, Complexity Accuracy Fluency (CAF), Oral Proficiency Interview (OPI)

CHAPTER 1: INTRODUCTION

In all educational domains, assessment play a crucial role in monitoring the learning progress and achievement of learners, as well as the effectiveness of teaching methods. As second language (L2) proficiency is transitional and invisible, assessment in second and foreign language learning is also necessary to regularly monitor learner progress. Lado (1961: p.25 – 29) proposed a model of language proficiency consisting of four "elements" (pronunciation, grammatical structure, lexicon, and cultural meaning) and four "skills" (listening, reading, writing, and speaking). He claims that these four skills' "degree of achievement" should be assessed separately as they develop at different speeds. In reality, however, L2 oral proficiency assessments are not administered as frequently as the other skills. This is due to the complexity of assessing productive skills (i.e., writing or speaking). While receptive skills (i.e., listening or reading) can be assessed easily by multiple-choice or fill-in-the-blank style items, productive skills require learners to provide constructive responses such as a composition or oral interaction with an examiner. Furthermore, written and spoken responses usually need to be rated subjectively by experts in the field. Although it is very necessary, it is unfortunately not practical to conduct frequent oral tests as it is too costly in terms of time, human resources, and money. One approach to the cost problem may be (semi-) automated scoring. The present study is a basic research study exploring algorithms that would support such automation.

In recent years, development and adaptation of automated scoring have become popular in the language testing industry (e.g., ETS, Pearson Education, etc.) (Xi, 2010). For example, the TOEFL iBT writing task is now rated in part by an automated scoring system. Enright and Quinlan (2010) claim that the consistency of computers can complement human raters' comprehensive and sophisticated judgments and increase the efficiency of scoring. For writing tasks, the computer looks at such surface linguistic features as length of essay and vocabulary sophistication to predict overall writing proficiency. It cannot replace humans since it does not score the essays with the same sophistication as trained human raters; however, when it is used to support human raters it can increase efficiency. This leads one to ask, could this be applied to speaking tests as well? Unfortunately, when compared to written responses, spoken discourse is much more fragmented, repetitive, and unstructured, which makes automated evaluation even more difficult (Xi, 2010). With current voice recognition technology, it is not yet feasible. In recent studies, researchers have attempted to tackle the problem by investigating the possibility of using fluency variables (e.g., how fast one speaks or how many repetitions one makes, etc.), quantifiable sub-components of speaking ability that do not require voice recognition, as predictors of overall speaking proficiency. The purpose of this study is to investigate relationships between objective fluency variables and oral proficiency, and then further examine a possibility of identifying a set of key fluency variables to predict overall oral proficiency, in combination with accuracy and complexity variables.

Generally speaking, the terms *oral proficiency* and *fluency* are used interchangeably to refer to the ability to orally use the second language (L2) at ease, as in “she speaks Italian fluently.” However, in linguistics, the meaning of *fluency* is more distinct. In classical theory, Skehan (1989) first proposed that *fluency* is one of the key elements of L2 proficiency, along with *complexity* (e.g., syntactic complexity, elaboration in speech, etc.), and *accuracy* (e.g., grammatical correctness.) The *Complexity, Accuracy,* and *Fluency* (CAF) indexes are said to be able to “expediently and reliably gauge proficiency in an L2” (Larsen-Freeman, 1978, p. 469), but the definitions and interpretations of these terms differ among researchers. Although researchers agree that fluency is indeed an important component of L2 oral proficiency, discussion on the definition of fluency has been on for almost a half century, and they have not yet to come to a consensus

(Chambers, 1997; Fillmore, 1979; Fulcher, 2003; Lennon, 1990; Schmidt, 1992). In an attempt to avoid confusion, many have argued for restricting the use of fluency to temporal characteristics of spoken discourse, because they are observable, quantifiable, and therefore reliable (Chambers, 1997; Leclercq, Edmonds, & Hilton, 2014; Segalowitz, 2010).

Möhle (1984) was the first to study L2 fluency objectively with temporal variables of fluency such as speech rate (the number of words or syllables articulated per minute), length and positioning of unfilled pauses, frequency and positioning of filled pauses, repetitions, and self-corrections. Soon afterwards, Lennon (1990) and many others have followed Möhle's example and adjusted the number of variables in an attempt to establish a set of variables that strongly correlate with L2 proficiency. Many seem to agree that speed related aspects of speech are correlated with the listeners' perception of what is considered to be 'fluent' (e.g., Ginther, Dimova & Yang, 2010; Hirotani, Matsumoto & Fukada, 2017; Houston, 2016; Iwashita, Brown, McNamara & O'Hagan, 2008; Kormos and Dénes, 2004; Lennon, 1990; Riggenbach, 1991).

While such research has contributed significantly to understanding L2 fluency, most of these studies were conducted with a small number of participants at a restricted range of proficiency levels. Also, many of the audio data used for the analyses are responses to a controlled task (e.g, a recall picture description task or retelling a story of a short video), rather than spontaneous speech. As some researchers have pointed out, when subjects are allowed to prepare before the speech, their fluency tends to improve (Greene & Capella, 1986). Although it is important to control for task variability, it is also necessary to investigate what information can be obtained from audio samples that are spontaneous and unprepared. Such data can provide information about learners' true ability to carry on a conversation in a natural discourse. Furthermore, as Préfontaine and Kormos (2015) point out, while there is a considerable amount of literature discussing fluency regarding English learners, little is known for languages other than English. It is therefore necessary to investigate if the previous findings on English cases are applicable to languages other than English, especially in Japanese where empirical data is scarce. In addition, while most previous studies have attempted to find a single fluency variable that is able to predict L2 proficiency levels, as far as the author is aware, none have attempted to find a composite fluency variable to predict learner proficiency. Since L2 proficiency is a complex ability to capture, it may be necessary to take multiple variables into account.

The aim of the current study is to contribute to the field of language assessment, oral proficiency, and fluency by investigating the relationship between objective fluency measures and L2 oral proficiency and identifying a set of fluency variables that would function as good predictors of L2 oral proficiency. Utilizing audio samples from the American Council on the Teaching of Foreign Languages' (ACTFL) well-established speaking test; i.e., the ACTFL Oral Proficiency Interview (OPI), this study analyzes spontaneous speech samples collected from L2 Japanese learners at a wide range of proficiency levels.

CHAPTER 2: RESEARCH BACKGROUND

This section first introduces an assessment (i.e., the OPI), a database, and software that are relevant to the present study, and then presents a review of the literature.

Research Data

Oral proficiency assessment

Perhaps the most widely used speaking test that directly assesses oral proficiency in a wide variety of languages is the ACTFL OPI. In the early 1980s, ACTFL created the first set of foreign language proficiency guidelines which was specifically designed for foreign language educators (ATCFL, 2012). Based on the guidelines, ACTFL created an interview-format assessment that can assess examinees' oral proficiency. Since then, the OPI has been considered the gold standard for assessing how well a person speaks a language. In general, the OPI is widely known as a reliable and valid test for assessing language proficiency (Dandonoli & Henning, 1990; Megnan, 1986; Surface and Dierdorff , 2003; Thompson, 1995; Thompson et al., 2016; Watanabe, 1998;).

What is the OPI?

It is a well-known standardized test that assesses one's oral proficiency through a 30-minute face-to-face or telephone interview with a trained examiner. Four major proficiency levels assessed by the OPI are novice, intermediate, advanced, and superior, each of which has three sublevels, except for superior, which has no sublevels. The OPI elicits examinees' utterances in a manner similar to real conversations, and they are rated holistically by at least two trained human raters.

The OPI rating procedure

All OPI levels are awarded using the following procedure. First, a certified interviewer gives a first rating, and then a second certified rater conducts a second rating. If the two raters award different levels, then a third rater conducts a third rating. If two different levels are assigned for an audio sample, the lower level will be awarded. If all the raters disagree, the recording will be labeled as “not ratable” and will be discarded.

Validity

Dandonoli and Henning (1990) examined the construct validity of the OPI and the use of the ACTFL Proficiency Guidelines. The data was obtained from 60 L2 French students and 59 ESL students, and four ESL and nine French certified OPI testers. Each student participated in the OPI, and other ACTFL tests including reading, listening, and writing. After conducting descriptive statistics, multitrait-multimethod construct validation, latent trait scalar analysis, and examining comparability of proficiency guidelines across language groups, the authors concluded that the results of the analyses provided “considerable support for the use of the guidelines as a foundation for the development of proficiency tests and for the reliability and validity of the OPI” (p. 20).

For the OPI in Japanese, Watanabe (1998) investigated the concurrent validity of the OPI by examining the relationship between awarded OPI levels and Japanese Proficiency Test[1] (JPT) scores. Although JPT does not have a speaking section, the JPT is designed to assess overall language ability of examinees, including speaking ability. It does so by predicting speaking ability from other abilities measured directly by the JPT such as language knowledge, reading,

[1] This is an older version of Japanese-Language Proficiency Test (JLPT).

and listening ability. The participants were 65 students learning Japanese at a U.S. university, enrolled in Second, Third, and Fourth year Japanese courses. The reported inter-rater reliability for the OPI rating was found to be high at (*r* = .96, *p* <.001). Correlation coefficients were computed between the awarded OPI levels and JPT scores (i.e., Listening Comprehension, Character Recognition, Reading Comprehension, and the total score). Positive correlations were found with the highest correlation coefficient of *r* = .71 (*p* <.001) with the total score. This result confirms that examinees who scored high on JPT also scored high on OPI and vice versa, which means the OPI can distinguish higher level examinees from lower level examinees well. The findings of this study support the validity of OPI in Japanese. One notable concern about this study is the use of JPT scores as comparison measures. The underlying assumption is that learners develop all skills evenly, but there has been a study demonstrating that it is not the case (Hirotani, Matsumoto, Fukada, 2017).

Reliability

Although OPI test-retest reliability reports are scarce compared to OPIc (i.e., computer-administrated equivalent of the OPI), the OPI is known to have high interrater reliability. Megnan (1986) investigated the interrater reliability of certified ACTFL OPI testers on a sample of 40 L2 French examinees. Megnan reported that a strong interrater agreement ($K = .72_{\text{Cohen's Kappa}}$) was found, and all disagreements in testers' ratings were within adjacent sublevel groups. Similarly, Thompson (1995) presented interrater reliability for the OPI under the 1986 guidelines. By analyzing 795 double-rated interviews, the interrater reliabilities (Pearson correlation) were calculated for French (*r* = .87), Spanish (*r* = .85), Russian (*r* = .90), English (*r* = .84) and German (*r* = .86). These results are evidence for the OPI's reliability across the languages.

Surface and Dierdorff (2003) have also provided a comprehensive analysis of the ACTFL OPI reliability by reporting interrater agreement and consistency across 19 languages, under the 2000 version of the guidelines. This study analyzed a total of 5881 interviews and their awarded scores by experienced ACTFL-certified testers. The languages included English, Mandarin, French, German, Italian, Russian, Spanish, Hebrew, Czech, Arabic, Vietnamese, Portuguese, Polish, Albanian, Hindi, Tagalog, Cantonese, Korean, and Japanese. A series of correlation coefficient tests revealed that 80 percent of the ratings across all 19 languages showed perfect agreement. Also, the consistency and rater agreement levels (i.e., Pearson correlation, Spearman rank-order correlation, Kendall's *tau*, Goodman-Kruskal *gamma,* and Cohen weighted *kappa* coefficient) between languages did not differ significantly between languages that are more commonly taught and less commonly taught (p. 512). Similarly to Thompson's (1995) research finding, any discrepancies found between the raters were within adjacent sublevels. When looking at Japanese data, there were 307 speech samples for Japanese OPI and the calculated interrater consistency values were all very high ($r = .981$, $R = .971$, $\tau = .933$, $\Gamma = .984$, $K_{wt} = .924$). These findings provide evidence for the ACTFL OPI's reliability.

The L2 Japanese learners' conversation database

The National Institute for Japanese Language and Linguistics (NINJAL) developed the "L2 Japanese learners' conversation database (https://db3.ninjal.ac.jp/kaiwa/)" in 2009. The database was developed for the purpose of investigating the general proficiency of learners of Japanese residing in Japan, and is open to the public. The database includes information about each participant's awarded OPI proficiency level, first language, age, gender, occupation, country of origin, length of stay in Japan, audio recording as well as a transcription of the OPI with a tester. There are 337 transcribed interviews available, and 215 of them are accompanied by audio recordings. The reason why there are fewer recordings than the transcriptions is that some participants did not consent to have their audio recordings released.

CAF Calculator

A few decades ago, a common approach for fluency studies was to transcribe audio data, count and calculate fluency measures manually by closely analyzing the speech samples and their transcribed data. It required a great amount of time and effort, and was one of the main reasons why the field of fluency studies was not as popular as others. Recently, with the technological advances, more and more tools have been created to lessen the burden for data analyses. The *CAF Calculator* (Fukada, Hirotani & Cantrell, 2019) is one such newly developed instruments. It is a free software program that can automatically calculate objective measures related to complexity, accuracy, and fluency. Audio must be annotated beforehand, but when it is used in combination with *Praat* (Boersma & Weenink, 2017), another piece of software which allows researchers to annotate audio files, and *Syllable Nuclei v2* (De Jong & Wempe, 2008), a Praat script software that can automatically detect and annotate syllables and sounding/silent portions of audio, it can greatly facilitate the calculation of fluency measures. The *CAF Calculator* was developed to promote more studies on speech production.

Literature Review

Complexity, Accuracy, and Fluency

In Second Language Acquisition (SLA), many researchers hold the view that L2 proficiency is understood as a multidimensional construct rather than a unitary one, and it can be captured by the concepts of complexity, accuracy, and fluency (CAF) (Housen, Kuiken & Vedder, 2012). The origin of CAF can be traced back to the 1970s (e.g. Hunt, 1965), when L2 researchers started applying the research findings for L1 grammatical complexity and accuracy to L2 research and found that these indexes can “expediently and reliably gauge proficiency in an L2” (Larsen-Freeman, 1978, p. 469). Later, Skehan (1996) proposed a L2 proficiency model that combined these three components together for the first time. The acronym CAF was then first introduced by a journal, Applied Linguistics, Special Issue “Complexity, Accuracy, and Fluency in second language acquisition research.” The definitions of the three components are still work in progress, but most commonly, complexity is characterized as “the ability to use a wide and varied range of sophisticated structures and vocab in the L2, accuracy as the ability to produce target-like and error-free language, and fluency as the ability to produce the L2 with native-like rapidity, pausing, hesitation, or reformulation” (Housen et al., 2012, p.2). Empirically, through factor analyses, these three components were identified as distinct and competing areas of L2 proficiency (e.g. Skehan & Foster, 1997), suggesting that all three components must be considered in L2 research together, not separately (Larsen-Freeman, 2009). Theoretically, the three components have been claimed to underlie the L2 developmental sequence (Skehan, 1998; 2003). It starts from internalization of new L2 elements leading to greater interlanguage (IL) system (i.e., greater complexity), followed by the modification of the IL systems or L2 structures (i.e., greater accuracy), and finally reach the level where these structures are automatized so that learners obtain greater control over their performance (i.e., fluency).

Various procedures have been employed to capture or evaluate CAF in applied linguistics including holistic/subjective, and objective quantitative measures, but the latter seems to be the preferred method in L2 production (Housen et al., 2012). One important issue that has been addressed by some researchers is the validity of those measures and their inconsistent application (Housen & Kuiken, 2009; Koizumi, 2005; Pallotti, 2009; Sakuragi, 2011). Sakuragi (2011) points out that some measures have been used inconsistently among researchers. Using a measure, "number of words per syntactic unit" as an example, Sakuragi (2011) explains that sometimes it is used as a syntactic complexity measure, and sometimes as a fluency measure. In addition, many of the CAF measures that were originally developed for L2 writing studies have been adapted to L2 speaking studies (e.g. Ortega, 1999; Skehan & Foster, 1999); however, some researchers such as Foster, Tonkyn, and Wigglesworth (2000) claim that some measures need adjustments to account for unique characteristics of speaking.

Sakuragi (2011), claiming the importance of validity of measurements, investigated the construct validity of 10 CAF measures using L2 Japanese speech samples. Narrative speech samples were collected from 113 university-level Japanese as a Second Language (JSL) students whose ACTFL OPI levels were intermediate to advanced. The following 10 CAF measures were calculated from the narrative production: syntactic complexity measures (i.e., number of clauses per Analysis of Speech [AS-unit], and number of subordinate clauses per AS-unit), lexical complexity measures (i.e., number of word types per 100 words, and the Guiraud index), accuracy measures (i.e., percentage of error-free AS-units, number of errors per AS-unit, and number of errors per clause), and fluency measures (number of words per minute, and number of dysfluency markers per minute). The result of a factor analysis revealed that the validity of syntactic complexity and that of accuracy measures were supported, but not lexical complexity and fluency measures. Sakuragi's (2011) result found that the fluency variables, the speed measure and the dysfluency measure did not share the same factor as one construct of fluency, and the author discussed it might be due to a lack of consensus or consistency of the definition of fluency, and its measurements. The author encourages more research to be conducted with a greater number of fluency measures to investigate the validity of fluency measures. The author also pointed out that although there is an extensive number of CAF investigations done in Indo-European languages, there are not enough studies done in other languages, such as Japanese.

Encouraged by Sakuragi's (2011) research, this study reviews definitions of fluency and investigate which fluency measures can validly capture L2 proficiency when combined with accuracy and complexity measures. Since the measures for accuracy and complexity were found to be valid in Sakuragi's (2011) study, the current study adopts one measure from each category, and include various fluency measures to investigate more about fluency.

Definitions of Fluency

Smooth delivery of a message plays an important role in successful oral communication. In general, the term *fluency* has extended meaning and is used interchangeably with the term *overall oral proficiency* to refer to L2 speakers' ability to use an L2 orally with ease (e.g., "she speaks English fluently"). However, in linguistics, the meaning of *fluency* is more distinct.

Fillmore (1979) was one of the first to study *fluency* in terms of L1. He conceptualized fluency in four different categories: "1) the ability to talk at length with few pauses, 2) the ability to talk in coherent, reasoned sentences, 3) the ability to have appropriate things to say in a wide range of contexts, and 4) the ability to be creative and imaginative in their language use such as to express their ideas in novel ways, or to create and build on metaphors" (p. 51). Although some of these characteristics of fluent speech of L1 speakers can be extended to the domain of L2 acquisition studies, some are problematic when applied without modifications to L2 learner speech.

Pawley and Syder (1983) provided one of the first definitions of L2 speakers' native-like fluency as "the native speaker's ability to produce fluent stretches of spontaneous connected discourse" (p. 191). However, as Chamber (1997) points out, there seems to be an overlap between first and second language domains where second language performance becomes as highly competent as ‘native-like’ speech. Nonetheless, Pawley and Syder’s definition narrowed down Fillmore’s broad definitions, and it has become a basis for later studies in L2.

Lennon (1990) argues that literature in EFL uses the term fluency in two senses: a broad, and a narrow sense. In the broad sense, fluency refers to global oral proficiency, while in the narrow sense it refers to an isolatable component of speech production such as speech tempo and smoothness, that is “native-like rapidity.” Citing Lennon’s article, Schmidt (1992) chose to restrict the use of the term to the narrow sense of fluency. For example, it is possible for one to speak an L2 with fluidity but with many errors. Schmidt explains that “such speaker is not fluent under the global proficiency definition but can be called fluent if we identify fluency with the processing of language in real time, rather than with language as the object of knowledge” (p. 358). Also, Schmidt added that he restricts his discussion of fluency to “the productive processes involved in the planning and delivery of speech,” rather than regarding fluency as the receptive processes. He argues that fluency in speech production is an “automatic procedural skill” (p.358), and he emphasizes that this procedural skill is a performance aspect rather than the knowledge, and it develops as more production processes become automatic.

From a cognitive approach, Skehan and Foster (1999) proposed that *fluency* is one key element of L2 oral proficiency, along with *complexity* (e.g., syntactic complexity, elaboration in speech, etc.), and *accuracy* (e.g., grammatical correctness.) Skehan and Foster defined *fluency* as “the capacity to use language in real time, to emphasize meanings, possibly drawing on more lexicalized systems” (p. 96), and argues that *fluency*, *complexity*, and *accuracy* are in a trade-off competition relationship for attentive resources of L2 speakers. They explain that because L2 learner’s attention is limited, their performance will be the result of negotiation and prioritization of one performance area over the others, depending on situations. For example, if one focuses more on accuracy when speaking, it might lead to a lack of fluency. Similarly to Schmidt's (1992) argument, Skehan and Foster also believe that fluent speakers do not require much effort because the psycholinguistic processes of speech planning and production are automatized.

Later, Skehan (2009) also focusing on the Lennon's narrow sense of fluency, further refined the definition by breaking it down to three categories of fluency: "1) breakdown (dys) fluency (i.e., index by pausing), 2) repair (dys)fluency (i.e., indexed by measures such as reformulation, repetition, false starts, and replacements), and 3) speed (i.e., with measures such as syllables per minute)." Since then, there has been an ongoing debate on the definition of fluency.

Although many researchers agree that fluency is a fundamental component of L2 oral proficiency, no consensus on what fluency is has been established (Chambers, 1997; Fillmore, 1979; Fulcher, 2003; Lennon, 1990; Schmidt, 1992). Since the broad sense's conceptualization of the term remains vague (Fulcher, 2003) and is hard to operationalize in real life, Lennon's narrower sense and the use of temporal measures of fluency (e.g., speech rate, pausing, and mean length of run) have attracted researchers' attention more. While some researchers argue that the complex nature of fluency cannot be reduced to a handful of temporal measures (Sajavaara & Lehtonen, 1978), other researchers have indicated that some temporal variables are correlated with overall oral proficiency (Kormos & Denes, 2004; Lennon, 1990; Riggenbach, 1991). Based on these findings, some researchers have argued for restricting the term of fluency to temporal and other fluency-related characteristics of spoken discourse, because those are observable, quantifiable, and therefore reliable (Chambers, 1997; Leclercq, Edmonds, & Hilton, 2014; Segalowitz, 2010).

From a cognitive linguistics perspective, Segalowitz (2010) introduces three domains of fluency: *cognitive fluency*, *utterance fluency*, and *perceived fluency*. Segalowitz defines each domain as follows:

1) *Cognitive fluency* – the efficiency of the operational process of speech planning and assembling functions, and of the integration and execution.

2) *Utterance fluency* – "the set of objectively determined timing, pausing/hesitation, and repair features of the utterance, reflecting the impact of the cognitive fluency on the underlying speech production processes" (p. 49).

3) *Perceived fluency* – "the inference listeners make about a speaker's cognitive fluency based on their perception of utterance fluency" (p. 48).

Figure 2.1 is a graphical representation of the three domains of fluency and their relationships.

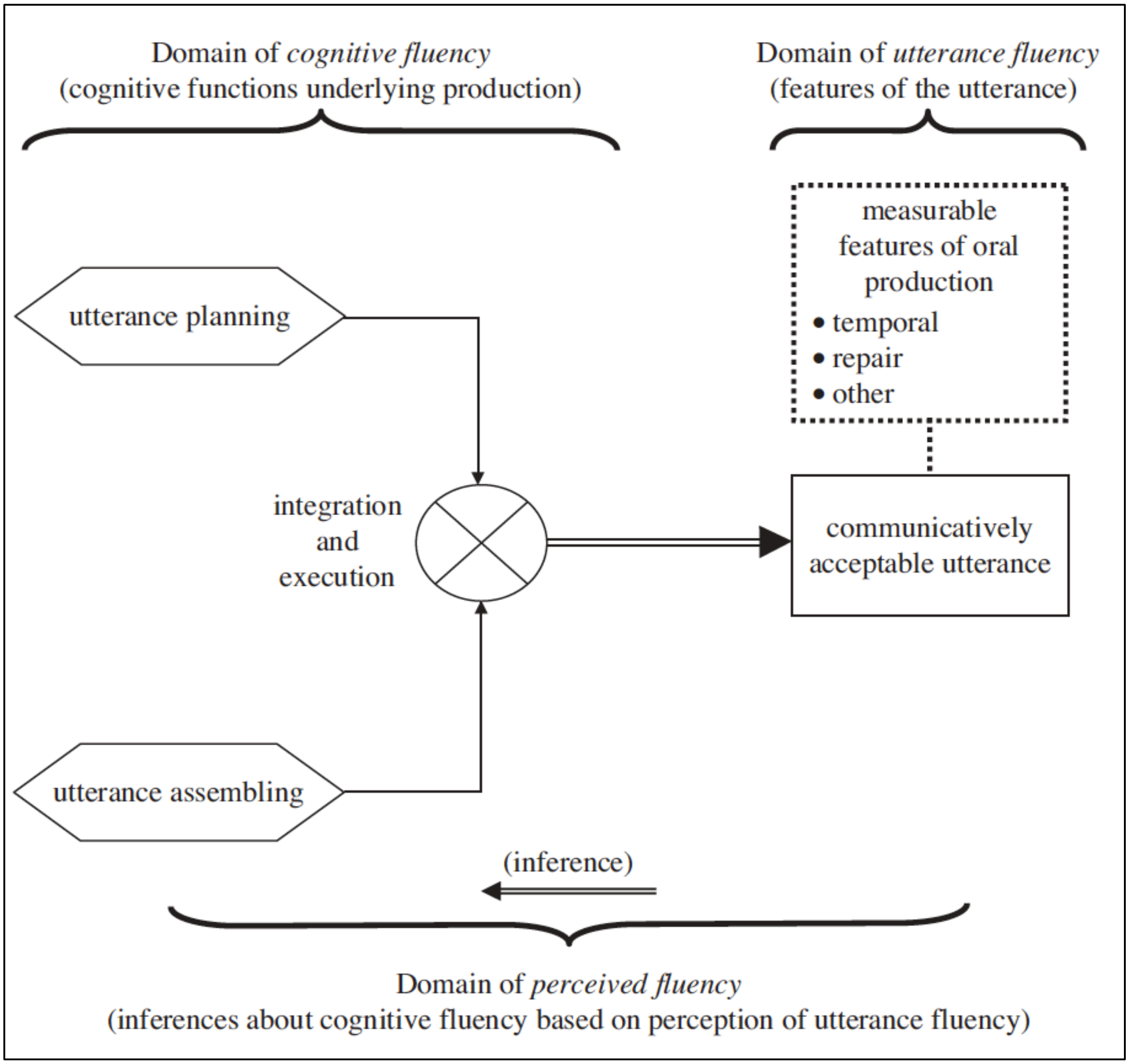
Domain of *cognitive fluency*
(cognitive functions underlying production)
Domain of *utterance fluency*
(features of the utterance)
measurable features of oral production
• temporal
• repair
• other
utterance planning
integration and execution
communicatively acceptable utterance
utterance assembling
(inference)
Domain of *perceived fluency*
(inferences about cognitive fluency based on perception of utterance fluency)

Figure 2.1 *Segalowitz's three domains of fluency (adapted from Segalowitz, 2010; p. 50)*

Segalowitz explains that cognitive fluency involves utterance planning and assembling, and these functions need to be integrated temporally to execute speech production with desired fluidity including timing, pausing/hesitation, and repair features. Utterance fluency is the objectively measurable features of oral production (e.g., timing, pausing/hesitation, and repair features), reflecting the impact of cognitive fluency (p. 49). Lastly, perceived fluency is the inference listeners make on how fluent (i.e., cognitive fluency) the speaker might be based on the observed utterance fluency. *Utterance fluency* is the key aspect that connects the three domains together. *Cognitive fluency* is reflected on *utterance fluency*, and some *utterance fluency* may cause listeners to judge some utterance as "communicatively unacceptable" (p. 51). Then, examining the relationship between *utterance fluency* and *perceived fluency* will reveal which objectively measurable feature of oral production has the greatest impact on listeners' judgment on how *fluent* the speaker is. For these reasons, this study adopts the concept of Segalowitz's (2010) three domains of fluency, and restrict the discussion of fluency to *utterance fluency.*

L2 Fluency Research Using Utterance Fluency

Möhle (1984) was the first to study L2 fluency objectively with temporal variables of fluency such as speech rate (the number of words or syllables articulated per minute), length and positioning of unfilled pauses, frequency and positioning of filled pauses, repetitions, and self-corrections. Soon afterwards, Lennon (1990) examined the fluency development of four English learners whose L1 was German using 12 fluency-related variables: eight temporal elements (e.g., speed of delivery, mean length run, and percentage of filled/unfilled pauses, etc.), and four dysfluency markers (i.e., repetitions, self-corrections, filled pauses, and percentage of repeated and self-corrected words). He found fluency improvements across three variables: speech rate, filled pauses per T-Unit (i.e., the smallest word group that could be considered a grammatical sentence), and percentage of T-Unit followed by pauses. Since then, many others have followed Möhle and Lennon's examples and adjusted the number of variables in an attempt to establish a set of variables that strongly correlate with L2 proficiency.

In order to investigate what features make a difference between highly fluent and nonfluent L2 speakers, Rigghenbach (1991) analyzed audio excerpts from six L1 Chinese non-native English speakers who were rated either "fluent" or "nonfluent" by 12 English instructors. The participants were asked to record a dialogue on an audiotape and submit it as a homework assignment of their ESL course. No specific topic for the dialogue was given for the purpose of obtaining reasonably "natural-sounding" conversation. Five-minute excerpts were then evaluated by 12 ESL instructors using a 7-point scale. Although the interrater reliability was not particularly high (r =.673), Rigghenbach justifies it is due to the nature of open-ended rating scale and the variability of how each rater interpreted the term "fluent." The participants were then divided into either 'fluent' or 'nonfluent' groups according to the judgment of the instructors, and their speech samples were analyzed for fluency-related variables. Rigghenbach used a total of 10 fluency related features including hesitation phenomena (i.e., micropause, hesitation, unfilled pause, filled pause), repair phenomena (i.e., retraced restart, unretraced restart), and rate and amount of speech (i.e., rate of speech, amount of speech, percentage of speech, total number of turns). She also analyzed seven interactive phenomena such as back-channeling, turn-taking, and so on as well; however, since those are not the focus of the current research, this section will only focus on the fluency-related measures. Table 2.1 summarizes Rigghenbach's fluency-related variables with definitions.

Table 2.1 *Riggenbach's (1991) Fluency-Related Variables*

Features	Variables	Explanations
Hesitation Phenomena	micropause	a silence of .2 seconds or less
	hesitation	a silence of .3 to .4 seconds
	unfilled pause	a silence of .5 seconds or greater
	filled pause	voiced "fillers," which do not normally contribute additional lexical information
	(a) nonlexical	fillers that are not recognized as words and that contain little or no semantic information
	(b) sound stretches	vowel elongations of .3 seconds or greater
	(c) lexical	fillers that are recognized as words but in context contribute little or no semantic information

Table 2.1 continued

Repair Phenomena	retraced restart	reformulations in which part of the original utterance is repeated (i.e., repetition, insertion)
	unretraced restart	reformulations in which the original utterance is rejected (= false start)
Rate and Amount of Speech	rate of speech*	number of words/sematic units per minute
	amount of speech	total number of words/sematic units (raw frequencies)
	percentage of speech	NNS to NS
	total number of turns	NNS and NS

The result of a Mann-Whitney U Test/Wilcoxon Rank Sum revealed significant differences between fluent group and nonfluent group in terms of rate of speech (*z*= 1.992, *p*= .046) and unfilled pause (*z*= 1.954, *p*= .049). Riggenbach concluded that “rate of speech and unfilled pauses contribute to judgment of nonfluency.” She also found that other factors such as repair phenomena seem to have less impact on listeners’ judgment of fluency, and therefore, the features 1) hesitation and repair, 2) rate of speech, and 3) interactive features are of unequal weight for potential predictors of fluency. Also, qualitative analysis of the speech suggested that more fluent speakers use more lexical fillers than nonfluent speakers, who tend to have more unfilled pauses or nonlexical fillers. However, she cautions that these results should be accepted with reservations due to the small sample size.

Towell, Hawkins, and Bazergui (1996) conducted a longitudinal study on 12 advanced learners of French, and compared their speech performance on the same task before and after their one-year-long study abroad experience. The participants were asked to watch a film and then were asked to re-tell the story individually. They also asked the same participants to provide the same re-telling of the story in their L1 (English) as well. They analyzed the recorded data for the following four temporal variables of fluency: speaking rate, phonation/time ratio, articulation rate, and mean length of run as summarized in Table 2.2.

Table 2.2 *Towell, et al.'s, (1996) Measured Temporal Variables*

Variables	Explanations/ Calculations
speaking rate	calculated by dividing the total number of syllables produced in a given speech sample by the amount of total time (including pause time) to produce the speech sample
phonation/time ratio	the percentage of time spent speaking as a percentage proportion of time taken to produce the speech sample
articulation rate	calculated by dividing the total number of syllables produced the amount of time taken to produce them, excluding pause time
mean length of run	calculated as the mean number of syllables produced in utterances between pauses of 28 seconds and above.

A series of *t*-tests were conducted to investigate the difference between time one (i.e., before the study abroad), time two (i.e., after the study abroad), and their L1 speech data. The results revealed that speaking rate, mean length of run, and articulation rate have all significantly increased after the study abroad ($t = 3.66, p < .01$; $t = 3.26, p < .01$; $t = 2.46, p <. 05$, respectively). They concluded that the increase on the speaking rate is mostly accounted for by changes in the mean length of run; therefore, they suggest that the best indicator of the development of fluency is mean length of run. The results also indicated that when a participant's speaking rate is high in their L1, it is likely that his/her L2 speaking rate will be high; however, the participants' performance in L2 at time two is still significantly slower than their L1 performance. Although they claim to have shown that the study abroad experience had a beneficial effect on the learners' fluency development, since there was no control group to compare with, the effect cannot be solely attributed to the study abroad experience. Also, the participants in this study were limited to one level: advanced learners. It is still not clear if these findings will be relevant to learners at different proficiency levels.

Kormos and Dénes (2004) examined audio samples from 16 English learners at two proficiency levels (advanced and low-intermediate) to investigate which fluency variables correlate with perceived fluency scores by six English teachers. Eight participants in the advanced group were Hungarian students studying English at a university, and the other eight participants in the lower-intermediate group were also Hungarian students studying English at a language school. Three of the judges were L1 Hungarian English teachers and the other three judges (i.e., teachers) were native speakers of English. In order to elicit speech samples, participants were asked to make up a story related to one of three cartoon strips of their choice. It is important to note that participants were provided with two minutes of planning time before the narrative task. The elicited speech samples were 2-3 minutes long on average. Kormos and Dénes explain that they chose an elicited narrative task for the following two reasons: 1) it is difficult to analyze speech phenomena in an interactive task, and 2) fixed content can eliminate the factor of varying cognitive loads on speakers. The speech samples were then rated by native and non-native English teachers intuitively using a 5-point scale, from least fluent to most fluent. The interrater reliability among the non-native teachers was $r = 0.78$, and native teachers was $r = 0.\ 73$. In this study, 10 temporal measures of fluency were employed to examine which variables predict native and non-native teachers' perception of fluency and distinguish non-fluent L2 speakers from fluent L2 speakers. Those measures and their definitions are summarized in Table 2.3.

Table 2.3 *Kormos and Denes's (2004) Examined Temporal Variables*

Variables	Explanations/ Calculations
speech rate	the total number of syllables produced in a given speech sample divided by the amount of total time required to produce the speech sample (including pause time greater than 3 seconds), expressed in seconds (Rigghenbach, 1991)

Table 2.3 continued

articulation rate	the total number of syllables produced in a given speech sample divided the amount of total time required to produce them in seconds (minus pause time), times 60
phonation-time ratio	the percentage of time spent speaking as a percentage proportion of the time taken to produce the speech sample (Towell et al., 1996, p. 91)
mean length of run	an average number of syllables produced in utterances between pauses of 0.25 seconds and above
the number of silent pauses per minute	pauses over 0.2 seconds were considered; total number of pauses divided by the total amount of time spent speaking expressed in seconds and then multiplied by 60
the mean length of pauses	the total length of pauses above 0.2 seconds by the total number of pauses above 0.2 seconds
the number of filled pauses per minute	the total number of filled pauses divided by the total amount of time expressed in seconds and was multiplied by 60
the number of disfluencies per minute	the total number of disfluencies (such as repetitions, restarts, and repairs) divided by the total amount of time expressed in seconds and was multiplied by 60
pace	the number of stressed words per minute
space	the proportion of stressed words to the total number of words

Spearman rank-order correlations were calculated to investigate the relationship between the temporal variables and the teachers' perceived fluency scores. Also, fluent and non-fluent speakers were compared by means of the Mann-Whitney U-test. Among the examined 10 temporal variables, the Mann-Whitney U-test revealed the significant differences between fluent and non-fluent speakers for the following five variables: speech rate (z = -3.04, p = .001), phonation time ratio (z = -2.31, p = .02), the mean length of run (z = -3.36, p = .001), the mean length of pauses (z = -1.99, p = .04), and pace (i.e., the number of stressed words per minute) (z = -3.36, p = .001). They reported that students with higher fluency scores spoke faster with fewer silent pauses, and produced longer stretches of discourse between pauses, used shorter pauses and uttered more stressed words within a minute than students with lower fluency scores (p.154). Also, fluent speakers can produce more accurate and lexically diverse speech than non-fluent speakers. From the results of the rank-order correlations, they concluded that "there is a set of variables that can predict the composite and the individual rater's fluency scores in a reliable way" (p. 154). Strong correlations were found between those variables and raters' fluency scores (NNS teachers and NS teachers, respectively) for the following variables: speech rate (r = .87, p < .001; r =.81, p < .001), phonation-time ratio (r = .80, p < .001; r = .74, p < .001), the mean length of run (r = .91, p < .001; r = .88, p < .001), the mean length of pauses (r = -0.58, p < .05; r = - 0.62, p < .001), and the number of stressed words per minute (i.e., pace) (r = .88, p < .001; r = .92, p < .001). They also noted that dysfluency phenomena (e.g., the number of filled and unfilled pauses) were not found to influence listener judgment. In conclusion, Kormos and Dénes claim that among these variables, speech rate, the mean length of run and pauses are the best predictors of fluency scores, regardless of the raters being NNS or NS.

Although this study has found the correlational relationships between some utterance fluency variables and perceived fluency, these findings were obtained from the speech samples of 16 students. Two other studies introduced previously also had a small sample size, six in Riggenbach (1991), and 12 in Towell et al. (1996). In order to generalize these findings, there is a need for studies with larger sample sizes (i.e. more than 30 to obtain stronger power in statistical analyses). Also, using "an intuitive 5-point fluency scale from least fluent to most fluent" as a subjective measure for comparison might not yield as accurate information as using a standardized and well-established oral proficiency test, such as the ACTFL OPI.

One such study that used a large sample size is Iwashita, Brown, McNamara, & O'Hagan (2008). They investigated the relationship between L2 English learners' speech features and their awarded holistic scores by trained raters. The examinees were English learners whose age, L1, length of study, and length of residence in an English-speaking country were varied, but were all studying English for the purpose of studying in the USA at the time of the data collection. A total of 200 audio samples collected through five tasks of TOEFL iBT were analyzed in terms of six features of fluency measures: filled pauses, unfilled pauses, repair, total pausing time, speech rate, and mean length of run. Those 200 audio files consisted of eight speech samples from each of five levels of TOEFL iBT for five tasks. The tasks used for this study were two independent tasks, where examinees were asked to express their opinions on a given topic without additional information, and three integrated tasks, where examinees were given additional information on a prompt either in reading or listening format and then asked to explain, describe, or recount the information. They were given 30 seconds to prepare and 60 seconds to speak for each task. This study examined the speech data in terms of *Linguistic Recourses* (i.e., grammatical accuracy, grammatical complexity, and vocabulary), *Phonology* (i.e., pronunciation, intonation, and rhythm), and *Fluency*, but this review will only focus on the *Fluency* analyses and results. Iwashita et al., examined the following six fluency measures (summarized in Table 2.4).

Table 2.4 *Iwashita et al.'s, (2008, p. 34) Examined Fluency Variables*

Variables	Explanations/ Calculations
filled pauses	calculated by the instances of ums and ers counted per seconds
unfilled pauses	calculated by counting the number of pauses of 1 second or more that occurred in the speech (Mehnert, 1998), per seconds
repair	calculated by the instances of repairs counter per seconds; (repairs refer to repetition of the exact words, syllables or phrases, replacement, reformulations, false starts, and partial repetition of a word or utterance (Freed, 2000)
total pausing time	calculated by adding up all the unfilled pauses and divided by the total speaking time
speech rate	calculated by dividing the total number of syllables produced in a given speech sample by the total time expressed in seconds (Ortega, 1999)
mean length of run	the mean number of syllables produced in utterances (Towell et al., 1996)

For the analysis, Iwashita et al, conducted an Analysis of Variance (ANOVA) with two factors (i.e. levels and task) to investigate which fluency features of speech performance distinguish among five levels of proficiency. ANOVA results revealed that *speech rate* (*F* (4, 189) = 71.32, *p* = .001, eta = .60), *number of unfilled pauses* (*F* (4,190) = 12.19, *p* = .001, eta = .20), and *total pause time* (*F* (4, 190) = 20.62, *p* = .001, eta = .30), have a clear relationship with proficiency levels. They further explained that these results indicate "higher-level learners spoke faster with less pausing, and fewer unfilled pauses" (p. 41). Additionally, they also concluded that *Vocabulary* (token[2]) and *Fluency* seem to have the greater influence on overall proficiency when compared to *Grammatical Accuracy* or *Pronunciation*.

These findings with the large sample size of 200 are convincing and they support the use of some key fluency measures in predicting L2 speakers' overall proficiency levels. However, they only examined five fluency variables and the relationships between other fluency variables and overall proficiency are not yet clear. It is more beneficial to include a variety of fluency variables in the analyses and explore which combinations of fluency variables can best predict overall proficiency.

Ginther, Dimova and Yang (2010) analyzed spoken responses of 150 L2 English speakers collected from the Oral English Proficiency Test (OEPT). The OEPT is a semi-direct test which was designed for screening the oral English proficiency of potential international teaching assistants at a university. The OEPT scale ranges from 3 to 6, and only those who earn a 5 or 6 are able to receive teaching assistantship positions at their university. Every year, approximately 500 examinees are tested, and of those, 30% are L1 Chinese speakers, 15% are L1 Korean speakers, 10% are Hindi speakers, and others. Since the range of proficiency levels of L1 Chinese speakers

[2] The number of words produced (Iwashita et al., 2008)

and Hindi speakers were known to be significantly different, 25 sets of responses from levels 3, 4, 5 from Chinese speakers, and 25 sets of responses from levels 5, 6 from Hindi speakers were selected randomly from a pool of OEPT responses. Also, a set of responses from 25 L1 English speakers is collected. This makes a total of 150 speech samples for the analysis. The speech samples were responses to an integrative task, where examinees were asked to express their opinions regarding a campus issue after listening to a narration. Examinees were given up to three minutes to prepare and two minutes to record their speech. The collected spoken responses were analyzed in terms of 15 temporal variables of fluency that can be categorized into the following three: "(1) the length of each pause in seconds, (2) the number of syllables uttered between pauses, and (3) the length of speech time in seconds between the pauses." The examined 15 temporal variables with explanations are summarized in Table 2.5 below.

Table 2.5 *Ginther et al.'s, (2010, p. 387) Examined Temporal Variables*

Variables	Explanations/ Calculations
Total Response Time	speaking + silent pause + filled pause time.
Speech Time	speaking time, excluding silent and filled pauses.
Speech Time Ratio	speech time/total response time
Number of Syllables	total number of syllables in a given speech sample was obtained to calculate mean syllables per run, speech rate, and articulation rate
Speech Rate	total number of syllables divided by the total response time in seconds. Total response time included both silent and filled pauses, and it was multiplied by 60.
Articulation Rate	total number of syllables divided by the sum of speech time and total filled pause time multiplied by 60. (i.e. articulation rate per minute)
Mean Syllables per Run	number of syllables divided by number of run in a given speech sample. Run were defined as number of syllables produced between two silent pauses. Silent pauses were considered pauses equal to or longer than 0.25 seconds.
Silent Pause Time	total time in seconds of all silent pauses in a given speech sample.
Number of Silent Pauses	total number of silent pauses per speech sample. Silent pauses were considered pauses of 0.25 seconds or longer.

Table 2.5 continued

Mean Silent Pause Time	silent pause time / number of silent pauses.
Silent Pause Ratio	silent pause time as a decimal percent of total response time.
Filled Pause Time	total time in seconds of all filled pauses in a given speech sample.
Number of Filled Pauses	total number of filled pauses.
Mean Filled Pause Time	filled pause time / number of filled pauses.
Filled Pause Ratio	filled pause time as a decimal percent of total response time.

As results of Spearman rank-order correlations, they found strong and moderately strong positive correlations between OEPT scores and speech rate (r= .72**), articulation rate (r= .61**), and mean syllables per run (r= .72**). They interpret these results to be “as examinees speak faster, say more, and pause less, their scores increase” (p. 388). Authors note that contrary to previous literature reporting the importance of filled pauses, they found no significant correlations between OEPT scores and any of the filled pause measures. Ginther et al. concluded that these findings confirmed theoretical expectations of the relationship between fluency measures and overall oral proficiency and suggested that these results support potential usage of these key temporal variables of fluency for automated scoring of overall oral proficiency. One small limitation of this study is that the audio samples used for this analysis were obtained from people who already had high English proficiency, because they must obtain at least 77 on TOEFL iBT in order to be accepted into graduate school at this university. Although these findings demonstrate the robustness of OEPT and of these key temporal measures by discriminating among examinees’ proficiency levels even within the restricted range, it is important to investigate the linearity of these measures in a wider range of oral proficiency levels.

Previous findings in the literature suggest that some key utterance fluency measures are correlated with overall oral proficiency and of those, some variables are more predictive of perceived fluency than others. However, most of these findings were yielded from speech samples of L2 English learners. Furthermore, these studies are insufficient for suggesting that previous findings can be transferred to other languages. More fluency studies in languages other than English are required.

Recently, Préfontaine, Kormos, and Johnson (2016) investigated the relationship between fluency measures and raters' perceptions of L2 fluency in French. The participants were 40 adult learners of French at various proficiency levels who were in an immersion context at a university in Québec. Their speech samples were elicited from three narrative tasks. Task 1 was a picture description task where participants were asked to describe six unrelated pictures. Task 2 was a story-retelling task where participants were asked to read a short passage about a horseback riding accident and then retell the story orally. Task 3 was a narrative task describing an 11-frame cartoon strip in sequence. For all tasks, participants were given three minutes for planning time, and there was no time-limit for providing a response. On average, the elicited speech samples were between three and four minutes. After speech samples were collected, 11 L1 French instructors rated them in terms of perceived CEFR levels (six levels from A1 to C2), pauses, and speed of speech. For rating pauses and speed of speech, two questions were asked using a six-point scale: 1) candidate can express themselves with few pauses/hesitations in French (1= a lot of hesitations, 6= very few hesitations), and 2) candidate can express themselves with reasonable speed in French (1= unreasonable speed, 6= very reasonable speed). Also, the speech samples were analyzed across four fluency variables: articulation rate, mean length of run, pause frequency, and average pause time. The examined fluency variables with explanations are summarized in Table 2.6.

Table 2.6 *Préfontaine et al.'s, (2016, p. 60) Examined Fluency Variables*

Features	Variables	Explanations/ Calculations
Speed fluency	Articulation rate	the total number of syllables divided by the total phonation time (excluding pauses) expressed in seconds.
Speed/Break down fluency	Mean length of run	the total number of syllables divided by the number of utterances between pauses of 0.25 seconds and above.
Breakdown fluency	Pause frequency	the total number of syllables divided by the total duration in seconds of the speech sample. (only including pauses above 0.25 seconds)
	Average pause time	the total duration of all pauses divided by the number of pauses in a given speech sample.

For statistical analyses, Préfontaine et al. conducted inter-correlations between utterance fluency measures, and found high inter-correlation levels (between r = .81 to .927) among articulation rate, mean length of run, and average pause time; however, pause frequency demonstrated weak correlations with articulation rate (between r = .233 and -.504), and moderately strong correlations with mean length of runs (between r = .233 and -.504). Authors explain that these high inter-correlations indicate that these three measures are all considered to represent the underlying construct of speed fluency, as opposed to pause frequency (i.e., breakdown fluency). Authors reported that one surprising finding from this result was that the correlation between articulation rate and pause frequency was negative, which suggests that speakers who pause less frequently make longer pauses. In addition, they also conducted a multiple regression analysis to see which of the four fluency variables are weighted more on each of the three perceived fluency variables. In their model, the independent variables were the four fluency variables and the dependent variables were perceived fluency ratings (i.e., CEFR level, speed, and pause). By comparing increased R-squared values, the results indicated that the relative importance of these four measures are in the order of mean length of runs (R^2= .324/.293/.289) [3]> Average pause time (R^2=

[3] The numbers represent effects of utterance fluency measures on CEFR/ pause/ and speed ratings, respectively.

.259/.241/.247) > Articulation rate (R^2= .225/.216/.229) > Pause frequency (R^2= .175/.166/.154). They reported that mean length of run was the most important predictor of perceived fluency, meaning that L2 speakers who can produce longer runs are likely to be perceived higher in fluency. Interestingly, their finding of the positive correlation between average pause time and perceived fluency rating suggested that speakers with longer average pause times were perceived to be more fluent in French. However, this study did not take the location of pauses into account, nor differentiated filled and unfilled pauses. Authors claim that this finding suggests that there is a prominent cross-linguistic variation specific to French, as opposed to the previous findings in ESL/EFL where longer pauses result in perceived dysfluency. This finding suggests that L2 fluency traits may be language specific. If that is the case, it is necessary to investigate if these differences are found in other languages as well.

To summarize, fluency variables that have been found to be good predictors of L2 oral proficiency in the literature are listed below.

1. *Speech Rate* (Ginther et al., 2010; Iwashita et al., 2008; Kormos & Dénes, 2004; Lennon, 1990; Rigghenbach, 1991; Towell et al., 1996)
2. *Mean Length of Run* (Ginther et al., 2010; Kormos & Dénes, 2004; Préfontaine et al., 2016; Towell et al., 1996)
3. *Articulation Rate* (Ginther et al., 2010; Kormos & Dénes, 2004; Préfontaine et al., 2016; Towell et al., 1996)
4. *Phonation Time Ratio* (Kormos & Dénes, 2004)
5. *Mean Length of Pauses* (Kormos & Dénes, 2004; Préfontaine et al., 2016)
6. *Unfilled Pauses* (Iwashita et al., 2008; Rigghenbach, 1991)
7. *Total Pause Time* (Iwashita et al., 2008)

8. *Amount of filled pauses per T-unit* (Lennon, 1990)

9. *Percent of T-units Followed by a Pause* (Lennon, 1990)

10. *Stressed Words per Minutes* (Kormos & Dénes, 2004)

Although there is some variability, many researchers seem to agree that speed related aspects of speech, especially *speech rate*, *mean length of run,* and *articulation rate* are strongly correlated with overall oral proficiency (Ginther, Dimova & Yang, 2010; Kormos & Dénes, 2004; Préfontaine et al., 2016; Riggenbach, 1991; Towell et al., 1996).

Using these observable and quantifiable constructs of speech performance, Chambers (1997) encourages more research into temporal variables in speech production to provide "concrete evidence which can contribute to a more precise definition of fluency" (p. 535). She further explains that since processes of language production are not accessible, studies of temporal variables in speech production "enable psycholinguistic research to gather valuable empirical evidence" (p. 538). In response to Chamber's suggestion, this study takes an empirical-based approach to capturing L2 oral proficiency with CAF measures (with a focus on fluency-related measures).

Fluency Studies in L2 Japanese

Previous fluency studies in Japanese can be divided into three types: (1) studies that investigated dysfluency factors from a pathological approach, (2) studies that investigated dysfluency factors of L2 speakers qualitatively, or (3) studies that investigated the correlation between listener judgment and fluency measures. Since the focus of this study is to investigate L2 Japanese fluency in the scope of second language acquisition using temporal measures of speech, this section will only review the third type.

In previous literature on L2 fluency in Japanese, there are some studies that have examined differences between NS and NNS of Japanese. For example, Ishizaki (2005) investigated differences in pausing patterns among NS of Japanese and beginner level L2 Japanese learners whose L1s were English, French, Chinese, and Korean. As a motivation for her study, she claims that English is quite different from Japanese in terms of syntax, rhythm, and accent, and these differences might affect speech patterns of the L2. The speech samples were collected from 10 native Japanese teachers, and 36 L2 Japanese speakers using read-aloud tasks. Of the speech samples, a 60-second strip was extracted from each sample for data analysis. There was no preparation time given before collecting data, and these learners had not been given any explicit instruction on pausing patterns in Japanese. The collected speech samples were analyzed in terms of mean length of run (i.e., the average length of speech between two pauses), frequency, length and positioning of pauses. From a series of ANOVA analyses, Ishizaki found significant differences between NS speech and NNS speech in the following features: 1) L2 learners' mean length of run is shorter than NS, 2) L2 speech has pauses at unnatural locations such as within a clause, and 3) L2 speech lack pauses at the end of sentences, or if they exist they are too short. Interestingly, Ishizaki reported that no significant difference was found among different L1 groups, which means that the characteristics of the L2 speech are similar regardless of L1. These results confirmed those of Ishizaki (2004). Ishizaki (2004) investigated the effects of pause patterns on NS listenability (i.e. perceived fluency) using a reading aloud task, recorded by both L1 and L2 Japanese speakers. The finding suggested that pausing location and length affect listeners' perception.

Houston (2016) investigated how L2 Japanese fluency develops through a language course at a university, by analyzing speech samples using fluency variables. This study collected speech samples from 30 novice level L2 Japanese students at the beginning and at the end of their school semester. The speech samples were elicited using two monologue tasks; a self-introduction task, and a typical school day task. They were asked to provide a short monologue about the given topic within 120 seconds. Data collection was conducted as part of their regular achievement tests within the course. The speech samples were analyzed in terms of 27 fluency-related variables included in the following categories: speech quantity, speed, pause, AS-unit related measures, and repair fluency. These 27 variables with explanations are summarized in Table 2.7.

Table 2.7 *Houston's (2016) Examined Fluency-Related Variables*

Features	Variables	Explanations/ Calculations
Speech quantity	Total response time	The time in seconds from the beginning of an audio response to the end of it
	Total number of syllables	All syllables in the file
	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 run) where a run is a sounding interval
	AS-Unit [4]speech rate	Effective syllable count / AS-Unit time * 60
Pause	Silent pause ratio	Silent pause time as a percentage of Total response time
	Silent pause count	The number of all silent pauses
	Silent pause time	The time in seconds of the sum of the duration of all silent pauses
	Filled pause count	The number of all filled pauses
	Filled pause time	The time in seconds of the sum 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 time in seconds of the sum of the duration of silent pauses falling within AS-Units
	Silent pause count between AS	The number of silent pauses between AS-Unit intervals

[4] = Analysis of Speech Unit (AS-Unit)

Table 2.7 continued

Pause	Silent pause time between AS	The time in seconds of the sum 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 time in seconds of the sum 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 time in seconds of the sum of the duration of filled pauses falling outside AS-Units
AS-Unit related measures	Number of AS-Units	
	Number of error free AS-Units	
	AS-Unit time	
Repair fluency	Repeat count	The number of repeat intervals (RP) on the DYSF tier
	Stutter count	The number of stutter intervals (ST) on the DYSF tier
	Self-correction count	The number of self-correction intervals (SC) on the DYSF tier

Of these 27 variables, speech quantity and speed related measures showed improvements from the beginning to the end of a semester. T-tests revealed that specifically *speech rate* (t = - 2.65, p < .05) and *mean length of run* (t = - 24.87, p < .05) showed significant improvement on the typical school day task. Houston claims that although pause related measures did not show significant improvements, these students produced more location appropriate pauses (i.e., in between clauses) at the second data collection when compared to the first collection. In Houston's (2016) study, she also investigated how these fluency-related measures are correlated to the NS judgment of overall oral proficiency. Using a rating rubric specifically developed for this study, NS instructors rated the same speech samples. They were recruited and trained to use the rubric to gain consistency in rating before the data collection. The participants' overall oral proficiency levels ranged from a score of 60 to 100. Correlation coefficients were calculated between awarded proficiency scores and their fluency-related variables. Of the 27 measures, *speech rate* (r = 0.60 task 1, r = 0.65 task 2) and *AS-unit speech rate* (r = 0.63 task 1, r = 0.70

task 2) showed strong correlation, and *mean length of run* also showed moderately strong correlation (*r* = 0.54 task 1, *r* = 0.50 task 2). These findings confirm previous findings in ESL/EFL studies; however, the oral proficiency level of the participants was restricted to beginner level. Studies with a wider range of oral proficiency levels are necessary to see if these findings can be applicable to students at other levels.

Recently, Hirotani, Matsumoto, and Fukada (2017) conducted a preliminary study to examine the relationship between L2 Japanese learners' proficiency test scores and their speech performance measured by fluency measures. The data was obtained from the International Corpus of Japanese as a Second Language (I-JAS), developed by the National Institute for Japanese Language and Linguistics, which contains speech samples from 215 L2 Japanese learners whose L1s are 12 different languages. The speech samples in the database were collected from six speaking tasks and six written tasks, along with informants' Japanese proficiency exam scores: the Japanese Computerized Adaptive Test (JCAT), and the Simple Proficiency-oriented Test (SPOT). This study extracted data of 15 L1 English Japanese learners' speech samples elicited by an interview task and a storytelling task, and their proficiency test scores (i.e. JCAT and SPOT) for the analyses. These informants' ages ranged between 18 and 24, and their proficiency test results showed that 14 out of 15 of them were rated as intermediate. Both JCAT and SPOT are widely known and used at institutions around the world to assess learners' language proficiency, and the correlation between them was found to be strong (Lee, et al., 2015). For the interview task, the informants were asked to participate in a 30-minute interview where they talked about various topics, which is in a similar format to the ACTFL's OPI. Within the 30-minute interviews, the researchers extracted a discussion part where informants were asked two questions: 1) "Which do you prefer, living in the city of in the country?" and 2) "Which is more important to you, time or money?" (p. 253). For the storytelling task, the informants were asked to tell a story based on four-frame and five-frame cartoons. They were asked to describe each scene in as much detail as possible. These speech samples were then analyzed in terms of two speed fluency measures (i.e. *speech rate* and *articulation rate*), and two breakdown fluency measures (i.e. *mean length of run* and *pause ratio*). Their explanations of the measures are summarized below in Table 2.8.

Table 2.8 *Hirotani et al.'s (2017) Examined Fluency Measures*

Variables	Explanations/ Calculations
Speech Rate	(Total number of morae[5]) / (Total response time) * 60
Articulation Rate	(Total number of morae) / (Speech time + Filled pause time) * 60
Mean Length Run	(Total number of morae) / (Number of run)
Pause Ratio	(silent pause time + filled pause time) / (Total response time) * 60

The correlation coefficients results found that there are strong correlations between JCAT score and all the fluency measures: *speech rate* ($r = .65_{\text{interview}}$, $r = .52_{\text{storytelling}}$), *articulation rate* ($r = .66_{\text{interview}}$, $r = .56_{\text{storytelling}}$), *mean length of run* ($r = .65_{\text{interview}}$, $r = .70_{\text{storytelling}}$), *pause ratio* ($r = -.60_{\text{interview}}$, $r = -.57_{\text{storytelling}}$). SPOT score on the other hand, showed slightly lower correlations compared to JCAT scores, especially with *speech rate* ($r = .47$) and *articulation rate* ($r = .45$) of the storytelling task. Of these four fluency measures, mean length of run on the storytelling task demonstrated strong positive relations with both JCAT ($r = .70$) and SPOT ($r = .66$) scores. The authors concluded that there is indeed strong relations between L2 Japanese learners' oral proficiency scores and the fluency measures, especially *speech rate* and *mean length of run*. In their analyses, they also investigated task differences between the less structured (i.e. interview task), and more structured (i.e. storytelling) task. They found that the fluency measures performed robustly across the tasks, but found slight differences on *speech rate,*

[5] Since Japanese is not a syllabic language, morae are used instead following previous studies (Ishizaki, 2004; 2005)

articulation rate, and *mean length of run* in terms of their correlations with the JCAT listening score. They explained that these differences are attributable to the nature of interview tasks where participants have to listen to what the interviewer is saying in order to respond, whereas there is no listening skills involved in the storytelling task. They found generally stronger correlation coefficients (ranging r = .60 to .66) in the interview task. The authors call for more studies on other tasks to see if other measures could perform even better, and they also claim that more studies are needed to investigate which measures could be combined to estimate learners' L2 proficiency more accurately. Encouraged by these findings, the current study investigates the relationship between L2 Japanese speakers' oral performance and their overall proficiency levels using CAF variables, and then a composite measure that can efficiently predict learners' L2 proficiency will be created. The current study also analyzes speech samples extracted from 30-minute interview sessions (i.e. OPI), on a descriptive task instead of a discussion task to see if different findings can be obtained.

Research Gaps and Motivation for The Present Study

While previous studies have contributed significantly to understanding the relationship between fluency measures and overall oral proficiency, there are some research gaps that need to be filled. With a few exceptions (e.g., Iwashita et al., 2008; Ginther et al., 2010), most of the previous studies' findings are yielded from a small number of subjects at a restricted range of proficiency levels. In order to obtain generalizable results, and also to capture a bigger picture of the relationship between fluency measures and oral proficiency, a larger sample size with a wider range of proficiency is needed. Also, sometimes the proficiency levels reported in the literature such as "advanced" or "intermediate" are vague terms, and not as clear as proficiency levels defined by well-established tests (e.g., TOEFL iBT, the ACTFL's OPI). The vagueness of the terms can raise a question about generalizability of research findings; hence it is ideal to use proficiency levels defined by well-established descriptors.

Moreover, most of the previous studies analyzed audio samples that were elicited by a controlled task (e.g, a recall picture description task or retelling a story of a short video), rather than spontaneous speech. As some researchers have pointed out, when subjects are allowed to prepare before the speech, their fluency tends to improve (Greece and Capella, 1986). Although it is important to control for task variability, it is also necessary to investigate what information can be obtained from audio samples that are spontaneous and unprepared. Such data can provide information about learners' true ability to carry on a conversation in a natural discourse.

Furthermore, as Préfontaine and Kormos (2015) point out, while there is a considerable amount of literature discussing fluency regarding English learners, little is known for languages other than English. As mentioned in the previous section, Préfontaine et al. (2016) found some language specific fluency phenomenon for L2 French speakers. If these differences can be found even among the Indo-European languages, it seems fair to expect some differences between English and other languages as well. It is therefore necessary to investigate if the previous findings in English cases are applicable to languages other than English, especially in Asian languages such as Japanese, where empirical data is scarce.

Taking these research gaps into consideration, the current study investigates the relationship between CAF (with a focus on fluency-related variables) variables and L2 oral proficiency. Utilizing audio samples from the ACTFL's well-established speaking test (i.e., the ACTFL OPI), this study analyzes spontaneous speech samples collected from 170 L2 Japanese learners at a wide range of proficiency levels. The first part is a correlational study, investigating the relationship between CAF variables and learner oral proficiency assessed by the ACTFL OPI.

The second part of the study created an optimal composite measure for predicting the OPI levels. The findings from this study will contribute to the discussion of using some key CAF measures as predictors of L2 overall proficiency, and future development of a (semi-) automated scoring system.

Research Questions

Research Question 1: Which CAF variables correlate with L2 Japanese proficiency levels measured by the ACTFL OPI, and to what extent do they correlate?

Research Question 2: Which combination of CAF variables can best predict examinees' L2 proficiency levels?

CHAPTER 3: METHODOLOGY

Overview

This chapter describes the research methodology for the current study. The chapter begins with a description of the speech samples used in this study, followed by the detailed information about the procedures of retrieving and coding of the data, and obtaining Complexity, Accuracy, and Fluency-related (CAF) measures. The chapter concludes with descriptions of the statistical analyses conducted to investigate the relationship between CAF measures and examinees' proficiency levels.

Data

The Database

The speech samples used in this study were obtained from the "L2 Japanese learners' conversation database (https://db3.ninjal.ac.jp/kaiwa/) ," which was published by the National Institute for Japanese Language and Linguistics (NINJAL) in 2009. The database includes 337 transcriptions of each participant's 30-minute face-to-face Oral Proficiency Interview (OPI) session, and 215 [6]of them have accompanied audio recordings of the OPI. The database also offers each participant's awarded OPI proficiency level, along with their background information (i.e., first language, age, gender, occupation, country of origin, length of stay in Japan).

The OPI samples were collected from 337 L2 Japanese learners residing in Japan, and their OPI levels ranged from Novice-Mid to Superior (i.e. nine proficiency levels). The data collection was conducted by a group of trained OPI testers in 2007, in six major cities: Tokyo,

[6] The number differs from the transcribed data simply because some informants did not consent to share their recordings.

Kyoto, Osaka, Nagoya, Kochi, and Kobe. The occupations of the participants are students at language schools (53%), university students (27%), businesspeople (4%), and others. Ages range from 10 to 39, and 66% were females while 31% were males. The major first languages of the participants in this database are Korean (53%), Chinese (17%), English (8%) and Other (22%). The length of stay in Japan varies from "3 – 6 months" (23%), "1 – 2 years" (23%), "2 – 3 years" (20%), "1 – 1.5 year" (17%), and "0.5 – 1 year" (16%).

Since this database can provide all the necessary data for the current study, it was selected as a suitable data source. For this study, 170 OPI audio recordings, transcriptions, the awarded OPI levels, and the informants' background information were extracted from the database.

The ACTFL OPI

For the purpose of the current study, a large number of speech samples from L2 Japanese learners of various proficiency levels were needed, and the proficiency levels must be determined by a reputable speaking test. The ACTFL OPI is widely known as a reliable and valid test for assessing language proficiency, and its scale has 11 levels, ranging from Novice to Distinguished, with sublevels[7]. Each speech sample is carefully rated by 2 to 3 human raters to determine the awarded OPI level. Moreover, the ACTFL OPI provides speech samples in a dialogue format rather than monologue format (e.g., talking to a computer), which is believed to have higher face validity for measuring test takers' functionality in daily conversations. The L2 Japanese conversation database provides the ACTFL OPI recordings collected from L2 Japanese learners at a wide range of proficiency levels.

[7] Note that there are no sublevels in *Superior* and *Distinguished* levels.

Participants

The participants in the current study were 170 L2 Japanese learners in Japan, whose OPI levels ranged from Novice-Mid to Superior. The occupations of these participants were students at language schools (42%), university students (35%), businesspeople (12%), and others (2%). Their ages ranged from 18 to 49, and 64% were females while 36% were males. The major first languages of the participants were Korean (37%), Chinese (22%), English (4%) and Other (28%). The length of stay in Japan varied from "1 – 6 months" (49%), "0.5 – 1 year" (20%), "1 – 2 years" (11%), "2 – 3 years" (4%), to "longer than 4 years" (16%).

Procedure

Retrieving Audio Recordings

For this study, 170 out of 215 audio recordings were retrieved from the database for analyses. Since the coding process is labor-intensive, the number of audio samples retrieved from Intermediate Mid and High were limited to 30 to ensure the quality of the coding process. The number of samples available in the database and the number of samples that were retrieved from each level are summarized below in Table 3.1.

Table 3.1 *Number of samples at each level*

OPI Levels	Available in the database	To be used in the study
Novice - Low	0	0
Novice - Mid	6	6
Novice - High	12	12
Intermediate - Low	21	21
Intermediate - Mid	58	30
Intermediate - High	47	30

Table 3.1 continued

Advanced - Low	27	27
Advanced – Mid	20	20
Advanced - High	19	19
Superior	5	5
Total	215	170

Selecting Audio Samples

Due to the adaptive nature of OPI, various speech tasks are given during a 30-minute interview session. Although there is a well-established procedure for administering the OPI and there are some key questions to be asked, the questions the testers ask are not predetermined. Rather, the tester asks questions according to the examinees' level and the natural conversation flow. That is quite common in real conversation and far more authentic than asking pre-determined questions that do not fit the conversation flow. Nonetheless, detailed criteria for sample selection must be established in order to minimize task variability.

Typical speech tasks that are included in OPI sessions are: answering yes-no questions, asking questions, telling a story, describing something or someone, expressing one's opinion about social issues, and talking about an abstract idea. This study focuses on the responses to descriptive speech tasks. In this study, a descriptive speech is defined as a speech segment where an examinee provides a new piece of information by explaining or describing a particular person, object, location, event, or one's thoughts or reasons. Example questions that elicit descriptive speech responses are "What are the differences between your hometown and where you live now?" or "Tell me about the most famous food from your city." If there were several descriptive task responses provided by one examinee, the longest response was selected (see Appendix I for example speech samples).

Also, to ensure the quality of the speech data, only the speech samples that meet all of the following criteria were selected:

1. The speech sample did not appear at the first or last 1 minute of the 30-minute OPI session.
2. The speech was continuous, single-turn, and not interrupted by the tester.
3. The examinee did not repeat the question.
4. The speech did not include any proper names[8].
5. The examinee's overall message was understood by the tester[9].
6. The examinee's speech provided a new piece of information beyond a simple one-word answer.
7. The topic of the speech was familiar and comfortable for the examinee to talk about[10].
8. The examinee did not use words from another language[11].

After undergoing the careful selection process, 170 speech samples were extracted from the original recordings and stored on the researcher's personal computer in .mp3 format. The average sample lengths are presented in Table 3.2 below.

Table 3.2 *The Average Sample Lengths*

OPI Levels	Average Length (seconds)
Novice - Mid	17.85
Novice - High	23.54
Intermediate - Low	34.53

[8] This is because the NINJAL database added data masking over proper names in order to protect sensitive personal information.

[9] This was determined by checking whether the tester asked a follow-up question for clarification or expressed confusion or not.

[10] For example, topics such as politics, visa status, sexual orientation, history can hinder examinees' speech.

[11] For example, a response "*Watashi no tsuma wa **housewife** desu*. (= My wife is a ***housewife***.)" would not be selected.

Table 3.2 continued

Intermediate - Mid	43.38
Intermediate - High	42.54
Advanced - Low	49.96
Advanced – Mid	52.11
Advanced - High	56.13
Superior	66.48

Data Processing

This section explains how the data was processed to obtain the complexity, accuracy, and fluency-related (CAF) measures. First, each speech sample was normalized, and noise reduced for the purpose of maximizing the quality of the audio. Moreover, the silent portion before and after the speech sample was eliminated in order to accurately calculate the speech time. A free audio editor and recorder, *Audacity® version 2.2.2.,* was used for this process. Next, the speech samples were annotated, using *Syllable Nuclei v2* (De Jong & Wempe, 2008) and *Praat* (Boersma & Weenink, 2017). Praat is a free computer software program that allows researchers to analyze audio data by transforming sound files into a sound wave and a spectrogram. It also allows researchers to make various custom annotations on the audio data for further analysis. Syllable Nuclei v2 is a free software script for Praat that can automatically detect syllables in running speech and annotate the sounding and silent parts of the speech. After the automated annotation was completed, the syllable counts, and the boundary locations were manually checked and corrected if necessary. Then, the researcher manually added other necessary annotations such as filled pause boundaries, AS-unit (i.e., Analysis of Speech Unit[12]) boundaries with and without grammatical errors, clause counts within an AS-unit, sound boundaries with dysfluency factors

[12] "a single speaker's utterance consisting of an independent clause or subclausal unit, together with any subordinate clause(s) associated with it" (Foster et al. 2000, p. 365).

(i.e., repetitions, short repetitions, and self-corrections), and sentence boundaries. The detailed definition of each annotation and coding criteria are introduced in the following section. For the purpose of checking the coding reliability, 10 out of 170 speech samples across the OPI levels were randomly selected, and then were sent to another expert in this field for the second coding. The detailed procedures and results of the Intraclass Correlation Coefficients (ICCs) calculations are presented and explained in the later section. Lastly, all annotated data was saved in .TextGrid format and then submitted to *CAF Calculator* (Fukada, et al. 2019), for computing CAF measures. A list of outcome measures with definitions and equations is presented in the later section for further explication.

Coding Scheme on Praat

As mentioned in the previous section, each speech sample was coded for the following 8 categories: (1) syllable count, (2) sounding and silent boundaries, (3) filled pause boundaries, (4) AS-unit boundaries, (5) AS-unit with or without grammatical errors, (6) clause counts within an AS-unit, and (7) sound boundaries for dysfluency factors (i.e., repetitions, short repetitions, and self-corrections), and (8) sentence count. The researcher carefully listened to each audio sample and annotates them on Praat screen. Figure 3.1 shows a sample screen of the coding process.

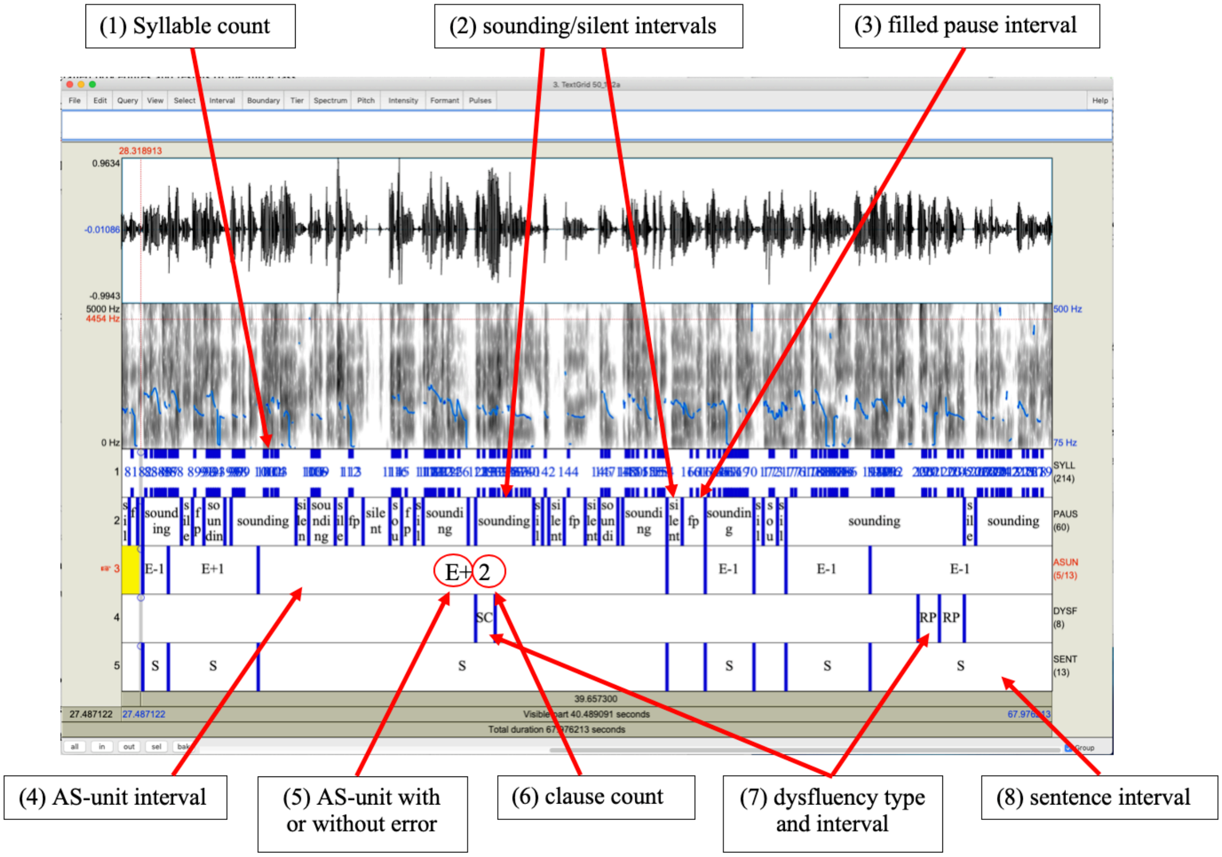

Figure 3.1 *Sample Screen of Coding Process*

For syllable count, the location and the number of syllables produced were annotated by the syllable markers, as shown in Figure 3.1. For categories 2, 3, 4, 7, and 8, the boundaries were marked to obtain intervals. All coding categories except for syllable count were annotated by unique symbols. The meaning of each symbol is summarized in Table 3.3 below.

Table 3.3 *Coding Symbols*

Coding Category	Annotation Symbol	Definition
2	sounding	sounding interval
2	silent	silent pause interval
3	fp	filled pause interval
4, 5	E+ / E-	AS-unit with / without errors
6	numbers	number of clauses in an AS-unit

Table 3.3 continued

7	RP	repeating interval
7	SC	self-correction interval
7	ST	short repetition interval
8	S	sentence interval

Term Definitions

Needless to say, consistent and reliable coding process cannot be achieved without clear definitions of the key terms. Table 3.4 is a summary of the key terms with explanations on how each term is defined for the current study.

Table 3.4 *Summary of the Key Terms*

Term	Definitions
syllable	The term 'syllable' in this study refers to the Japanese 'mora' unit. (Mora is a smaller unit than a syllable (Tamaoka & Terao, 2004)). This is because Japanese is a moraic language and most of the Japanese alphabets correspond to a mora rather than a syllable. The term 'syllable' is not changed to 'morae' in this study, for the purpose of the easier comparison to the previous studies where most of them are based on syllabic languages.
silent pause	In this study, the term 'silent pause' is defined as the silent portion of a speech that does not contain any utterance. A silence of 0.25 seconds or above is considered as one silent pause in this study, because it is the most widely employed detection criterion in the previous studies (Ginther et al., 2010; Houston, 2016; Park, 2016; Préfontaine et al., 2016;).
filled pause	Filled pause' in this study is defined as a portion of utterance that includes voiced fillers such as hmm, er, or um equivalents in Japanese. Japanese filler examples include the followings: あー, ああ, えー, えっと, うーん, なんか , なんだろう, まぁ. In addition to the above examples, some utterance where the speaker says "うん (=yeah)" or "はい (=yes)" to confirm what s/he has just said, are also considered a type of fillers.

Table 3.4 continued

AS-unit	The Analysis of Speech Unit (AS-unit) is a term proposed by Foster et al. (2000) and defined as "An AS-unit is a single speaker's utterance consisting of an independent clause, or sub-clausal unit, together with any subordinate clause(s) associated with either" (p.365). This study carefully follows the definitions and guidelines for detecting AS-units; however, some modifications are made in order to compensate for the unique difference between English and Japanese. One modification is that when two or more clauses connected with the TE-form of verbs (roughly equivalent to "and") have different subjects, the clauses are counted as two separate AS-units instead of one (Sakuragi, 2011).
error	'Error' in this study is limited to the grammatical errors and obvious vocabulary errors. Typical examples of grammatical error include mistakes on tense, conjugation, particles, conjunction, and so on. An example of vocabulary error is illustrated below: E.g., "私の町の有名な食べ物は、スカイです。フルーツのスカイ。" (= The most famous food in my hometown is *_Sukai_[13]. _Sukai_, the fruit.)
clause	The definition of 'clause' in this study was adapted from Hirotani (2009). Hirotani (2009) defines a clause as "an utterance with a predicate" (p. 422), and "an utterance including coordinate conjunctions such as te-, tari-, and shi-forms 'and', is considered a clause because such expressions contain predicates" (p.422). E.g., "ご飯を食べたら、お皿を洗います。" (= Once I'm done eating, I will do dishes.) In this case, there are two clauses within a sentence because there are two predicates "*taberu* (= to eat)" and "*arau* (=to wash)."
repetition	"A repetition is where the speaker repeats previously produced speech" (p. 368), following the definition of Foster et al. (2000).
Short repetition	A short repetition is defined as a portion of speech where the speaker continues producing unintentional repetition of sounds, especially the initial sound of a word.
sentence	A string of words that is complete in itself, typically containing a subject and predicate, and consisting of a main clause and sometimes one or more subordinate clauses. Usually Japanese sentences end with polite predicate suffixes (i.e., -desu and -masu) as an indicator of the end of a sentence; however, in spontaneous speech, sometimes those ending indicators are dropped and a falling tone on the ending predicate is used instead. In this study, that kind of falling tone is also considered as an indication for the sentence ending.

[13] The speaker actually means "watermelon," but she says *_waterlemon_ instead.

Coding Reliability Check

In order to test for coding consistency, 10 speech samples are randomly selected from the 170 samples, and then annotated by another expert in this field for second coding. The second coder carefully followed the same coding procedure described in the previous section. Once the coding is done, the annotated data is submitted to *CAF Calculator* to obtain the 48 CAF measures. The outcome measures obtained from the researcher and the second coder are compared. To assess the coding agreement between the researcher and the second coder, Intraclass Correlation Coefficients (ICCs) with two-way mixed effects model are computed using IBM SPSS Statistics version 25.

First, the three fluency measures that were found to be good predictors of L2 oral proficiency in previous literature, Speech Rate, Articulation Rate, and Mean Length of Run, are examined. For Speech Rate, average measure ICC is .993 with 95% Confidence Interval (C.I.) of .972 - .998. For Articulation Rate and Mean Length of Run, the ICCs are .980 (95% C.I. = .925 - .995) and .928 (95% C.I. = .647 - .983), respectively. Since the ICCs for all three measures are greater than .900, the results indicate that the coding of the researcher and the second coder are very similar.

Secondly, some base CAF measures are compared. Although there are 48 CAF measures, some of them are combination measures based on the 28 fundamental measures. Therefore, the following section reports the ICCs for the 28 measures. As shown in Table 3.5, the ICCs for 21 out of 28 measures (75%) are above .90. Repeat count (ICC = .6233) and self-correction time (ICC = .686) show the lowest ICCs; however, this result is somewhat expected since the occurrence of these measures are typically low and if there is disagreement, it affects the ICC results greatly. Nevertheless, the overall high ICCs support that the researcher's coding is reliable and consistent. Table 3.5.

Table 3.5 *Summary of ICCs*

CAF Measures	Intraclass Correlation Coefficient (95% C.I.)
Speech Time	.990 (.942 - .998)
Total Number of Syllables	.997 (.987 - .999)
Number of AS-units	.974 (.882 - .994)
Number of error-free AS-units	.850 (.320 - .964)
Number of AS-units with errors	.981 (.881 - .996)
Number of sentences	.984 (.936 - .996)
Clause count	.995 (.982 - .999)
Silent pause count	.967 (.869 - .992)
Silent pause time	.977 (.902 - .994)
Filled pause count	.934 (.676 - .984)
Filled pause time	.949 (.683 - .988)
Silent pause count within AS-unit	.944 (.786 - .986)
Silent pause time within AS-unit	.979 (.906 - .995)
Silent pause count between AS-unit	.945 (.779 - .986)
Silent pause time between AS-unit	.941 (.760 - .985)
Filled pause count within AS-unit	.946 (.779 - .986)
Filled pause time within AS-unit	.959 (.843 - .990)
Filled pause count between AS-units	.791 (.240 - .947)
Filled pause time between AS-units	.828 (.349 - .957)
Repeat count	.623 (.000 - .900)
Sutter count	.920 (.693 - .980)
Self-correction count	.727 (.000 - .933)
Repeat time	.811 (.190 - .954)
Short repetition time	.900 (.584 - .975)
Self-correction time	.686 (.000 - .923)
Effective syllable count	.982 (.930 - .995)
AS-unit time	.998 (.994 – 1.00)
Sounding count	.975 (.900 - .994)

Complexity, Accuracy, and Fluency-Related variables

Once all speech samples are coded and its coding reliability is assured, it is time to submit the data to the *CAF Calculator*. Although the CAF Calculator automatically computes 50 measures as outcome, for this study, some adjustments are made from the default output to obtain 48 CAF measures to be examined in this study. Table 3.6 shows a summary of the 48 measures and their explanation and calculations.

Table 3.6 *CAF Variables to Be Examined*

Category	**Sub-category**	**Measures**	**Explanation / Calculations**
Speech Quantity		Total response time	The time in seconds from the beginning of an audio response to the end of it
		Speech time	The sum of all sounding intervals in seconds (excluding fillers)
		Total number of syllables	All syllables in the file
		Effective syllable count	(Total number of syllables) – (syllables in repeat, short repetition, and self-correction intervals) – (syllables in fillers)
		Sounding count	The number of all sounding intervals
		Number of sentences	
		Number of AS-Units	
		AS-Unit time	The time in seconds of the sum of the duration of all AS-Units
		Phonation time ratio	(Speech time) / (Total response time) * 100
Speed		Speech rate	(Total number of syllables) / (Total response time) * 60
		Articulation rate	(Total number of syllables) / (Speech time + Filled pause time) *60
	Speech density	Mean length run	(Total number of syllables) / (Number of runs) where a run is a sounding interval
		Effective speech rate	(Total number of effective syllable) / (Total response time) * 60
		Effective articulation rate	(Total number of effective syllable) / (Speech time – DYSF time) * 60
Pause		Silent pause count	The number of all silent pauses
		Silent pause time	The time in seconds of the sum of the duration of all silent pauses
		Filled pause count	The number of all filled pauses
		Filled pause time	The time in seconds of the sum of the duration of all filled pauses
		Silent pause ratio	Silent pause time as a percentage of Total response time
		Silent and filled pause ratio	(Silent pause time + Filled pause time) / (Total response time) * 100

Table 3.6 continued

Pause		Silent pause count	The number of all silent pauses
		Silent pause time	The time in seconds of the sum of the duration of all silent pauses
		Filled pause count	The number of all filled pauses
		Filled pause time	The time in seconds of the sum of the duration of all filled pauses
		Silent pause ratio	Silent pause time as a percentage of Total response time
		Silent and filled pause ratio	(Silent pause time + Filled pause time) / (Total response time) * 100
	Pause location	Silent pause count within AS	The number of silent pauses within AS-Unit intervals
		Silent pause time within AS	The time in seconds of the sum 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 time in seconds of the sum 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 time in seconds of the sum 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 time in seconds of the sum of the duration of filled pauses falling outside AS-Units
		Silent pause ratio within AS	(Silent pause time within AS-unit)/(Total response time)*100
		Silent and filled pause ratio within AS	(Silent pause time within AS-Unit + Filled pause time within AS-Unit)/(Total response time)*100
		Ratio of silent pause time between AS to total response time	(Silent pause time between AS-unit)/(Total response time)*100
		Ratio of silent and filled pause time between AS to total response time	(Silent pause time between AS-unit + Filled pause time between AS-unit)/(Total response time)*100

Table 3.6 continued

Repair (dysfluency)		Repeat count	The number of repeat intervals (RP) on the DYSF tier
		Short repetition count	The number of short repetition intervals (ST) on the DYSF tier
		Self-correction count	The number of self-correction intervals (SC) on the DYSF tier
		Repeat time	The total duration of repeat intervals (RP) on the DYSF tier
		Short repetition time	The total duration of short repetition intervals (ST) on the DYSF tier
		Self-correction time	The total duration of self-correction intervals (SC) on the DYSF tier
		DYSF time	The total duration of all dysfluency intervals on the DYSF tier
		Repeat ratio	(Repeat time) / (Total Response Time) * 60
		Short repetition ratio	(Short repetition time) / (Total Response Time) * 60
		Self-correction ratio	(Self-correction time) / (Total Response Time) * 60
		DYSF ratio	(DYSF time) / (Total Response Time) * 60
Complexity		Clause count	The number of clauses in file
		Syntactic complexity	Clause count / Number of AS-Units
Accuracy		Number of error-free AS-Units	
		Number of AS-units with errors	
		Error-free AS-unit ratio	(Number of error-free AS-Units) / (Number of error-free AS-Units + Number of AS-Units with errors) * 100

As the purpose of this study is to (1) examine the relationship between the CAF measures and L2 proficiency levels, and (2) find the most efficient combination of CAF measures for predicting the proficiency levels, this study includes a variety of fluency-related measures and some simple measures for quantifying complexity and accuracy of speech. The fluency-related measures can be classified into four major categories: speech quantity, speech speed, pause, and repair.

As previously mentioned, the most frequently reported fluency variables that are positively correlated to the proficiency levels in literature are: speech rate, articulation rate, and mean length of run. All three of them are speed related measures; however, the interpretation and calculation of each measure is slightly different. Ginther et al. (2010) explains that speech rate is "the most general and inclusive measure" (p.382), whereas articulation rate "focuses on the amount of time required for a speaker to physically produce speech" (p.382), Ginther et al. (2010) further explains that although mean length of run is typically regarded as part of speech related measures, it is better described as "a density measure that appears to represent both syntactic well-formedness and vocabulary" (p. 382 – 383). This concept of "speech density" is of great interest to the current study; however, the researcher found the variable, mean length of run, to be little problematic. Mean length of run is typically calculated in the following formula: (total number of syllables) / (total number of runs in a given speech sample). What is problematic here is that this formula does not take repairs and other dysfluency phenomena into account. For example, if the total number of syllables include some repetitions, those syllables should not be counted because that portion does not add density to the speech. If one wants to capture speech density more accurately, it makes more sense to eliminate syllables for dysfluency markers such as short repetition, self-correction, and repetition, from the formula. Therefore, in this study, two new speech speed variables are proposed: effective speech rate, and effective articulation rate (refer to Table 3.4 for calculations). “Effective syllable” focuses only on meaningful production of a speaker that does not include any syllables used for fillers or dysfluency markers. Both variables can account for speech speed and speech density at the same time. Effective speech rate represents how fast a speaker can produce effective syllables within the total response time, and effective articulation rate represents how fast a speaker can produce effective syllables if s/he is not interrupted by any pauses or dysfluency markers.

In addition, there is a subcategory within pause related measures. This is because the researcher is interested if a pause location has any relationship with proficiency levels. For pause locations, within and between AS-units are considered.

Data Analysis

For statistical analyses IBM SPSS Statistics version 25 is used. The independent variable is the ACTFL OPI levels and the dependent variables are the CAF measures. To address the Research Question 1, Spearman’s rank-order correlation coefficients are calculated between CAF measure and the OPI levels, as well as the correlation among the CAF measures. Then, a series of Multiple Linear Regressions is conducted to investigate which combination of the CAF variables can best predict L2 Japanese speakers’ proficiency levels (Research Question 2).

CHAPTER 4: RESULTS AND DISCUSSION FOR RESEARCH QUESTION 1

Introduction

This chapter describes and discusses the results of the data analyses to investigate the primary research question in this study: Which CAF variables correlate with L2 Japanese proficiency levels measured by the ACTFL OPI, and to what extent do they correlate? The chapter begins with the descriptive statistics of the 48 CAF measures and OPI levels, followed by the results and discussion of Spearman's rank order correlation between the 18 CAF measures and OPI levels. The number of CAF measures are reduced to 18 measures that do not get affected by the length of extracted speech samples for this analysis. The third section presents the results and discussion of further analysis on Spearman's rank order correlation coefficients among 18 CAF measures. The chapter concludes by summarizing the research findings in response to the primary research question.

Descriptive statistics of the 48 CAF measures

This section presents descriptive statistics of the variables to be examined. Table 4.1. displays means, standard deviations (*SD*), minimum and maximum values, and skewness for all the variables. Overall, most variables are more or less normally distributed, except for the following five variables: 22. SILENT PAUSE TIME WITHIN AS-UNIT, 28. FILLED PAUSE TIME BETWEEN AS-UNIT, 37. SHORT REPETITION TIME*,* 40. REPEAT RATIO, and 42. SELF-CORRECTION RATIO. Their skewness values are emphasized in bold in Table 4.1. They all display slight positive skewness; however, it is expected. First of all, they are all time-related variables and time does not allow negative values. Although the occurrence time for dysfluency phenomena tend to be short, some people may take much longer time, which then pulls the distribution to the positive side. Since this is a natural phenomenon, it is safe to conclude that all variables are suitable for the analysis.

Table 4.1 *Descriptive Statistics of the 48 CAF Measures and the OPI levels*

Variable	*n*	*M*	*SD*	Min.	Max.	Skew
Oral Proficiency						
OPI Levels	170	5.07	2.00	1.00	9.00	-0.02
Speech Quantity						
1. Total response time	170	44.01	22.20	6.40	123.95	1.01
2. Speech time	170	27.18	14.75	2.66	93.28	1.01
3. Total number of syllables	170	171.24	106.86	15.00	553.00	1.27
4. Effective syllable count	170	156.29	103.11	9.00	551.00	1.37
5. Sounding count	170	19.42	9.68	3.00	52.00	0.82
6. Number of sentences	170	4.12	2.36	0.00	16.00	1.42
7. Number of AS-units	170	4.48	2.44	1.00	17.00	1.48
8. AS-Unit time	170	38.86	20.18	6.40	118.60	1.11
9. Phonation time ratio	170	61.22	12.38	27.65	83.86	-0.40
Speed						
10. Speech rate	170	228.94	71.57	88.31	437.32	0.41
11. Articulation rate	170	321.68	69.16	171.92	515.39	0.25
Speed / Density						
12. Mean length run	170	8.88	3.67	3.33	23.80	1.27
13. Effective speech rate	170	207.80	75.73	53.33	419.64	0.40
14. Effective articulation rate	170	366.86	64.74	239.36	574.99	0.60
Pause						
15. Silent pause count	170	20.04	11.38	2.00	72.00	1.40
16. Silent pause time	170	12.66	8.51	2.43	50.56	1.78
17. Filled pause count	170	8.74	6.25	0.00	33.00	1.15
18. Filled pause time	170	4.02	3.23	0.00	17.14	1.41
19. Silent pause ratio	170	29.39	11.08	8.77	60.81	0.42
20. Silent & filled pause ratio	170	38.37	12.24	14.54	72.35	0.41
Pause Location						
21. Silent pause count within AS	170	15.49	9.51	1.00	66.00	1.78
22. Silent pause time within AS	170	9.16	6.63	0.25	47.20	**2.11**
23. Silent pause count between AS	170	4.55	3.37	0.00	19.00	1.49
24. Silent pause time between AS	170	3.50	3.08	0.00	17.83	1.68
25. Filled pause count within AS	170	6.38	5.33	0.00	29.00	1.50
26. Filled pause time within AS	170	2.78	2.62	0.00	14.30	1.78
27. Filled pause count between AS	170	2.36	2.14	0.00	13.00	1.62
28. Filled pause time between AS	170	1.24	1.38	0.00	10.30	**3.03**
29. Silent pause ratio within AS	170	21.38	10.11	1.07	53.74	0.91

Table 4.1 continued

Variable	*n*	*M*	*SD*	Min.	Max.	Skew
Pause Location						
30. Silent & filled pause ratio within AS	170	27.45	11.68	1.07	72.35	0.95
31. Ratio of silent pause time between AS to total response time	170	8.01	5.94	0.00	32.77	1.37
32. Ratio of silent & filled pause time between AS to total response	170	10.92	7.66	0.00	42.05	1.32
Dysfluency						
33. Repeat count	170	1.63	1.72	0.00	9.00	1.71
34. Short repetition count	170	1.89	1.97	0.00	9.00	1.35
35. Self-correction count	170	0.97	1.29	0.00	6.00	1.43
36. Repeat time	170	1.02	1.18	0.00	6.54	1.86
37. Short repetition time	170	0.63	0.79	0.00	4.24	**2.04**
38. Self-correction time	170	0.76	1.10	0.00	5.61	1.84
39. DYSF time	170	2.41	2.25	0.00	11.47	1.56
40. Repeat ratio	170	1.54	1.87	0.00	11.90	**2.12**
41. Short repetition ratio	170	0.84	1.01	0.00	6.04	1.96
42. Self-correction ratio	170	1.03	1.73	0.00	11.80	**3.09**
43. DYSF ratio	170	3.41	2.88	0.00	17.34	1.57
Complexity						
44. clause count	170	10.54	6.81	0.00	39.00	1.03
45. syntactic complexity	170	2.44	1.35	0.00	9.00	1.48
Accuracy						
46. number of error-free AS-Units	170	2.67	2.26	0.00	13.00	1.40
47. number of AS-units with errors	170	1.81	1.47	0.00	8.00	1.29
48. error-free AS-unit ratio	170	54.44	31.24	0.00	100.00	-0.22

Relationships between the 18 CAF measures and OPI levels

In order to understand relationships between CAF measures and the OPI levels, Spearman's rank-order correlation coefficients were calculated among the 18 CAF measures and OPI levels. Spearman's rank-order correlation was selected as a statistical method instead of Pearson correlation because the OPI levels are ordinal data, and while Pearson correlation only detects linear relationships, Spearman's rank-order correlation can provide insight into non-linear relationships as well. For the analyses, the number of CAF measures were reduced to 18 because some of the 48 measures are dependent of the length of speech samples. Since this study extracted single-turn descriptive speech samples from 30-minute OPI sessions, the length of extracted samples varied. The selected 18 CAF measures are not affected by the length of extracted speech samples. The removed measures are all base measures that appear in the formulas of the 18 measures. In this sense, the selected 18 measures are representative of the removed measures.

Table 4.2. shows the correlation coefficients of the 18 CAF variables against OPI levels. This section focuses on the relationship between each CAF variable and OPI levels. Then, the next section reports on how those 18 CAF variables are correlated to each other.

Table 4.2 *Rank-order Correlation Coefficients for 18 CAF measures vs. OPI levels*

Category	Sub-category	Measures	Correlation Coefficient
Speech Quantity		9. Phonation time ratio	.57**
Speed		10. Speech rate	**.74****
		11. Articulation rate	**.66****
	Speech Density	12. Mean length run	**.67****
		13. Effective speech rate	**.78****
		14. Effective articulation rate	**.67****
Pause		19. Silent pause ratio	-.55**
		20. Silent and filled pause ratio	-.56**
	Pause Location	29. Silent pause ratio within AS	-.44**
		30. Silent and filled pause ratio within AS	-.37**
		31. Ratio of silent pause time between AS to total response time	-.24**
		32. Ratio of silent and filled pause time between AS to total response time	-.24**
Dysfluency		40. Repeat ratio	-.41**
		41. Short repetition ratio	.02
		42. Self-correction ratio	.04
		43. DYSF ratio	-.42**
Complexity		45. Syntactic complexity	**.63****
Accuracy		48. Error-free AS-unit ratio	.46**

* Indicates *p*. < 0.05., ** Indicates *p*. < 0.01. Bold letters indicate strong correlation.

Results

Strong correlations[14] in the range of |*r*|= .60 – .79 are observed for the following six measures in three categories:

1. *Speed*: 10. SPEECH RATE (*r*= .74**), 11. ARTICULATION RATE (*r*= .64**)
2. *Speed/Density*: 12. MEAN LENGTH RUN (*r*= .67**), 13. EFFECTIVE SPEECH RATE (*r*= .78**), 14. EFFECTIVE ARTICULATION RATE (*r*= .67**)
3. *Complexity*: 45. SYNTACTIC COMPLEXITY (*r*= .63**).

According to the results, all *Speed*-related variables demonstrated strong correlations with the OPI levels. Within the *Speed* category, 10. SPEECH RATE demonstrated higher correlation than that of 11. ARTICULATION RATE. For the *Speed/Density* category, 13. EFFECTIVE SPEECH RATE showed the highest correlation coefficient than other two. Also, 45. SYNTACTIC COMPLEXITY from the *Complexity* category was found to be strongly correlated to the OPI levels. Interpretation of these findings is discussed later in this section.

Moderately strong correlations in the range of |*r*|= .40 - .59 were observed for the following seven measures in five categories.

1. *Speech Quantity*: 9. PHONATION TIME RATIO (*r* = .57**)
2. *Pause:* 19. SILENT PAUSE RATIO (*r* = -.55**), 20. SILENT & FILLED PAUSE RATIO (*r* = -.56**)
3. *Pause Location*: 29. SILENT PAUSE RATIO WITHIN AS (*r* = -.44**)
4. *Dysfluency*: 40. REPEAT RATIO (*r* = -.41**), 43. DYSF RATIO (*r* = -.42**)
5. *Accuracy*: 48. ERROR-FREE AS-UNIT RATIO (*r* = .46**).

Among these measures, *Speech Quantity* and *Pause* variables showed stronger correlation than that of *Pause Location*, *Dysfluency*, and *Accuracy* measures. Although 40. REPEAT RATIO and 43.

[14] Measures with strong correlations are emphasized in bold in Table 4.2.

DYSF RATIO showed moderately strong correlations, non-significant correlations were found for 41. SHORT REPETITION RATIO (*r* = .02) and 42. SELF-CORRECTION RATIO (*r* = .04) .

Weak correlations in the range of |*r*|=.20 - .39 were found for the following three measures in the *Pause Location* category: 30. SILENT & FILLED PAUSE RATIO WITHIN AS (*r* = -.37**), 31. RATIO OF SILENT PAUSE TIME BETWEEN AS TO TOTAL RESPONSE TIME (*r* = -.24**), 32. RATIO OF SILENT & FILLED PAUSE TIME BETWEEN AS TO TOTAL RESPONSE (*r* = -.24**). Interpretations of these findings is discussed in the following section.

Discussion

From the results, strong correlations were found with *Speed* (*Speed/Density*) and *Complexity* measures. Of all six categories, *Speed* measures demonstrate the strongest relationship with the OPI levels, because all five measures in this category show strong correlation coefficients. This finding is in align with the literature, because the top three most frequently reported measures correlating strongly to the OPI levels (i.e., SPEECH RATE, MEAN LENGTH OF RUN, ARTICULATION RATE) are also found to be strong in this study. Among these three measures, SPEECH RATE has the strongest correlation, followed by MEAN LENGTH OF RUN and ARTICULATION RATE, in that order. SPEECH RATE showing the strongest relationship with the OPI levels confirms the previous findings in the literature (e.g., Ginther et al., 2010; Iwashita et al., 2008; Kormos & Dénes, 2004). Interestingly, although most of the previous findings are based on English, the results of the current study suggest that they can be extended to Japanese as well. Then, it indicates that these speed related measures, especially SPEECH RATE, may be applicable cross-linguistically.

Though SPEECH RATE shows the highest correlation among the three measures (i.e., SPEECH RATE, MEAN LENGTH OF RUN, ARTICULATION RATE), this study includes two new measures that represent *Speed* and *Speech Density* (i.e., EFFECTIVE SPEECH RATE, and EFFECTIVE ARTICULATION

RATE). When correlation coefficients are compared, both new measures show stronger correlations than their original counterparts (i.e., SPEECH RATE, and ARTICULATION RATE). Notably, EFFECTIVE SPEECH RATE shows the highest correlation coefficients (*r*= .78**) of all *Speed* related measures. This is probably because while SPEECH RATE only accounts for speech speed, EFFECTIVE SPEECH RATE takes speech density (only counting meaningful production) into account. This means that as OPI level advances, their rate of producing meaningful syllables increases.

Other than *Speed* related measures, it is interesting to see that the variable representing *Complexity* (i.e., SYNTACTIC COMPLEXITY) also shows correlation as high as *Speed* related variables. Although the complexity measure in this study is a very simple one, the result indicates that *Complexity* has as strong a relationship with the OPI levels as *Speed* measures.

Moderately strong correlations were found in *Speech Quantity, Pause* (also *Pause Location*), *Dysfluency*, and *Accuracy* variables. The results suggest that as the OPI rating advances, the amount of time spent on speech increases as pausing time decreases. This finding supports the previous findings (Iwashita et al., 2008; Reggenbach, 1991). Also, the results revealed that there is only a weak relationship between *Pause Location* and the OPI levels, except for SILENT PAUSE RATIO WITHIN AS-UNIT. It means that pause location does not matter much unless it is silent pauses occurring within AS-units. The negative correlation coefficient suggests that the more silent pauses within AS-units there are, the lower their OPI rating is.

For *Dysfluency* variables, among the three dysfluency types (i.e., repetition, short repetition, and self-correction), only REPEAT RATIO shows a moderate relationship with the OPI levels. The negative correlation coefficient suggests that as examinees' OPI level advances, their amount of time spent for repetition decreases; however, the occurrence of short repetition or self-correction does not have much effect on the OPI levels.

The *Accuracy* variable is also found to be moderately correlated. Similarly to the *Complexity* variable, the *Accuracy* measure follows a simple method of quantifying grammatical/vocabulary accuracy by counting the number of AS-units with or without errors. Despite the simple method, it still demonstrated moderate relationship with the OPI levels. The positive correlation coefficient of ERROR-FREE AS-UNIT RATIO suggests that as oral proficiency improves, examinees can speak more accurately in terms of grammar and vocabulary. This finding, that all *complexity, accuracy*, and *fluency* variables demonstrated moderate to strong correlations to the OPI levels supports the literature that CAF are indeed important components of L2 oral proficiency (e.g., Larsen-Freeman, 2009; Skehan, 1996).

Summary

In summary, *Speed/Density* and *Complexity* variables are strongly correlated with the OPI levels. Also, *Speech Quantity, Pause*, *Pause Location* (i.e., SILENT PAUSE RATIO WITHIN AS-UNIT only), *Dysfluency* (i.e., REPEAT RATIO only), and *Accuracy* are found to be moderately correlated with the OPI levels. The results indicate that as examinees' oral proficiency improves, they can speak faster, can produce more meaningful syllables, with more complex and accurate sentences that do not contain frequent repetitions or pauses (especially silent pauses within AS-unit).

Relationships among the 18 CAF measures

The previous section explored the relationships between each of the 18 CAF measures and OPI levels; however, some measures are closely related to each other. In order to understand the relationships among the CAF measures, this section focuses on the intercorrelation among the 18 measures.

Table 4.3 *Correlation Matrix of 18 CAF measures*

	9	10	11	12	13	14	19	20	29	30	31	32	40	41	42	43	45	48
9. Phonation time ratio	1.00																	
10. Speech rate	**.87****	1.00																
11. Articulation rate	.62**	**.88****	1.00															
12. Mean length run	**.85****	**.89****	.75**	1.00														
13. Effective speech rate	**.83****	**.97****	**.87****	**.88****	1.00													
14. Effective articulation rate	.41**	**.78****	**.88****	.60**	.79**	1.00												
19. Silent pause ratio	**-.86****	-.74**	-.39**	-.74**	-.70**	-.35**	1.00											
20. Silent&Filled pause ratio	**-.99****	**-.86****	-.62**	**-.85****	**-.82****	-.42**	**.87****	1.00										
29. Silent pause ratio within AS	-.74**	-.65**	-.37**	-.68**	-.63**	-.32**	**.85****	.74**	1.00									
30. Silent&Filled pause ratio within AS	-.76**	-.65**	-.47**	-.70**	-.64**	-.29**	.68**	.76**	**.90****	1.00								
31. Ratio of silent pause time between AS to total	-.33**	-.26**	-0.09	-.21**	-.22**	-0.12	.43**	.35**	0.00	-.17*	1.00							
32. Ratio of Silent&Filled pause time between AS	-.37**	-.31**	-.20**	-.24**	-.28**	-.16*	.34**	.38**	-0.08	-.19*	**.94****	1.00						
40. Repeat ratio	-.25**	-.31**	-.32**	-.29**	-.43**	-.26**	.16*	.24**	0.14	.20**	0.09	0.11	1.00					
41. Short repetition ratio	0.05	0.01	-0.05	-0.03	-0.06	-0.06	-0.11	-0.04	-0.12	-0.07	0.00	0.05	-0.06	1.00				
42. Self-correction ratio	0.03	0.00	-0.04	0.00	-0.13	-0.10	-0.09	-0.04	-0.09	-0.05	0.01	0.01	0.07	.26**	1.00			
43. DYSF ratio	-.18*	-.28**	-.30**	-.28**	-.47**	-.31**	0.08	.17*	0.07	0.12	0.03	0.06	.66**	.38**	.59**	1.00		
45. Syntactic complexity	.45**	.50**	.41**	.52**	.53**	.36**	-.42**	-.43**	-.23**	-.18*	-.36**	-.37**	-.28**	0.08	0.06	-.27**	1.00	
48. Error-free AS-unit ratio	.33**	.39**	.43**	.42**	.44**	.35**	-.24**	-.31**	-.34**	-.38**	0.12	0.10	-.24**	0.01	-0.07	-.30**	0.14	1.00

* Indicates significance at p<.05, **Indicates significance at p<.01,

Bold letters indicate strong correlation.

Results and Discussion

Table 4.3 reports a correlation matrix of the 18 CAF measures. As expected, some CAF measures are closely related to each other. Strong to very strong correlations are found especially among *Speech Quantity*, *Speed*, and *Pause* related measures. This analysis will focus on the high correlation[15] (above $|r|= .80$) as they might represent the same construct and are highly dependent, and those with weak correlations are considered as unique or relatively independent as a measure.

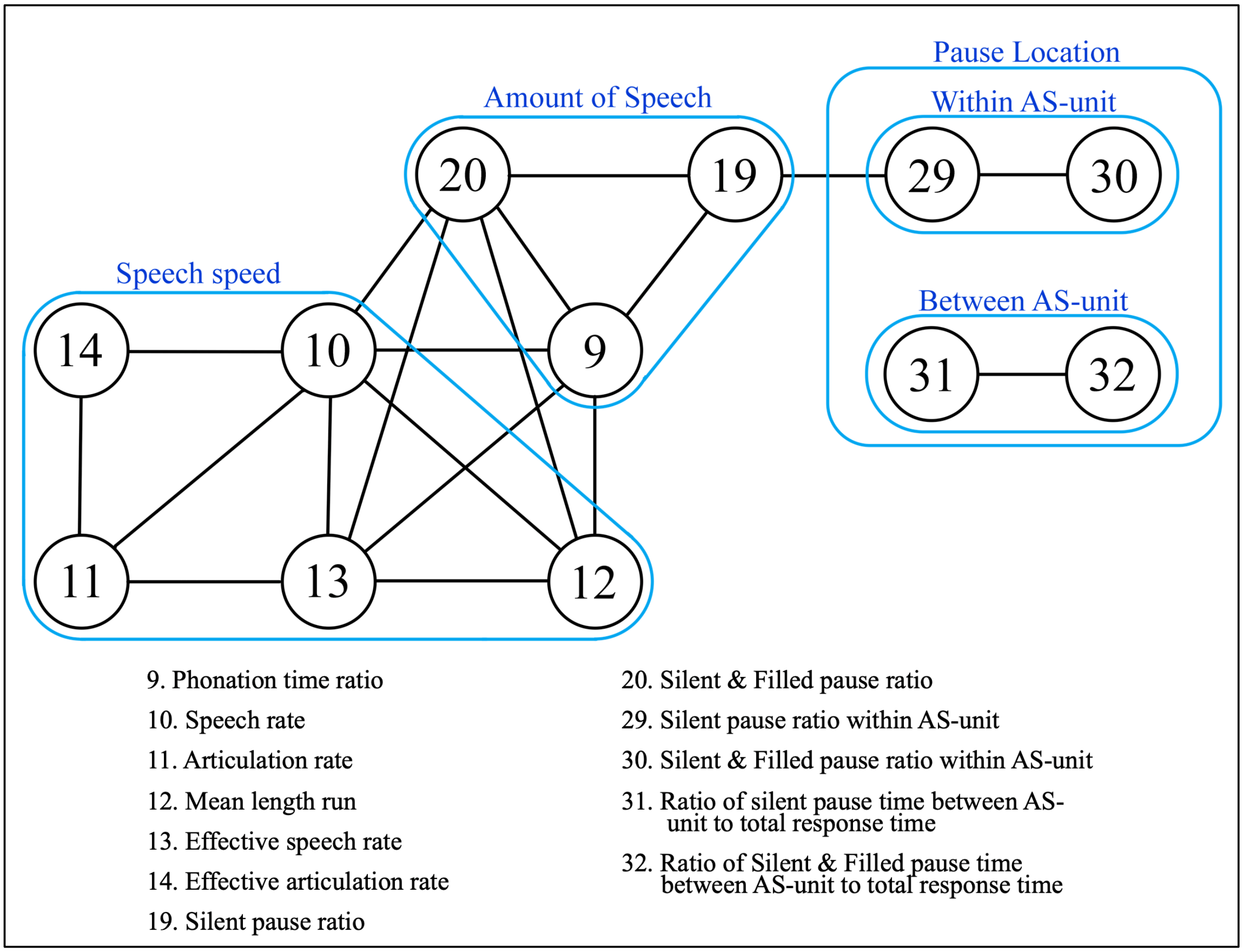
Pause Location
Amount of Speech
Within AS-unit
20
19
29
30
Speech speed
Between AS-unit
14
10
9
31
32
11
13
12
9. Phonation time ratio
10. Speech rate
11. Articulation rate
12. Mean length run
13. Effective speech rate
14. Effective articulation rate
19. Silent pause ratio
20. Silent & Filled pause ratio
29. Silent pause ratio within AS-unit
30. Silent & Filled pause ratio within AS-unit
31. Ratio of silent pause time between AS-unit to total response time
32. Ratio of Silent & Filled pause time between AS-unit to total response time

Figure 4.1 *Visualization Map for Very Strongly Correlated Variables*

[15] The very strong correlation coefficients are represented by bold letters in Table 4.3.

Figure 4.1 lists all pairs that are very strongly correlated with each other. The variables are then grouped together to visualize the relationships among them. As Figure 4.1 shows, the measures are complexly related to each other; however, they can be categorized into three major groups: *Speech Speed, Amount of Speech,* and *Pause Location*. There are five variables that represent *Speech Speed*: SPEECH RATE, ARTICULATION RATE, MEAN LENGTH OF RUN, EFFECTIVE SPEECH RATE, and EFFECTIVE ARTICULATION RATE. These results are expected because the only difference among them is whether they include or exclude pauses and dysfluency phenomena.

Amount of Speech category consists of PHONATION TIME RATIO, SILENT PAUSE RATIO, and SILENT & FILLED PAUSE RATIO. Although the latter two are pause-related measures, they also represent the amount of speech because the opposite of time spent for pausing is speaking time with or without filled pauses. Hence, these three variables are categorized together to represent the amount of speech.

Lastly, variables that represent *Pause Location* also show very strong correlations. What is interesting here is that only *Pause within AS-unit* variables are connected to the other variables and *Pause between AS-unit* variables are separated from the rest. By revisiting Table 4.3, stronger correlations are found between *Pause within AS-unit* variables and other fluency variables, than *Pause between AS-unit.* This finding suggests that while pauses made within AS-unit have greater impact on the other fluency variables, pauses between AS-unit do not.

Summary

This section examined the relationships among the 18 CAF measures. Some of the measures were found to be strongly correlated, especially among *Speech Speed, Amount of Speech,* and *Pause Location* variables. Within the *Pause Location* category, it seems that pauses occurring within AS-units have stronger relationship with the rest of the variables when compared to pauses that occur between AS-units.

Response to Research Question 1

This section answers Research Question 1 “Which CAF variables correlate with L2 Japanese proficiency levels measured by the ACTFL OPI, and to what extent do they correlate?” Results revealed that *Speed/Density*, and *Complexity* variables are strongly correlated (|*r*|= .60 – .79) with L2 Japanese OPI. The highest correlation coefficient was found for EFFECTIVE SPEECH RATE (*r* = .78), and the second highest was SPEECH RATE (*r* = .74), suggesting that *Speed*-related measures have the strongest relationship with the OPI levels. Also, *Speech Quantity, Pause*, *Pause Location* (i.e., SILENT PAUSE RATIO WITHIN AS-UNIT), *Dysfluency* (i.e., REPEAT RATIO), and *Accuracy* are found to be moderately correlated with the OPI levels. The results indicate that as examinees’ oral proficiency advances, they can speak faster, can speak more with meaningful syllables, as well as more complex and accurate sentences that do not contain frequent repetitions or pauses (especially silent pauses within AS-unit). In addition, the results revealed that among the CAF measures, variables representing *Speech Speed* (i.e., SPEECH RATE, ARTICULATION RATE, MEAN LENGTH OF RUN, EFFECTIVE SPEECH RATE, and EFFECTIVE ARTICULATION RATE), *Amount of Speech* (i.e., PHONATION TIME RATIO, SILENT PAUSE RATIO, and SILENT & FILLED PAUSE RATIO), and *Pause Location* (i.e., PAUSE LOCATION WITHIN AS-UNIT, and PAUSE BETWEEN AS-UNIT) are very strongly correlated to each other within each group.

CHAPTER 5: RESULTS AND DISCUSSION FOR RESEARCH QUESTION 2

Introduction

This chapter reports on several analyses that were conducted to investigate the second research question in this study: Which combination of CAF variables can best predict examinees' L2 oral proficiency levels? This study is exploratory and empirical-based. For the analyses, a series of Multiple Linear Regression analyses (MLR) was conducted to find a parsimonious model that can best predict the OPI levels in the most efficient way, given the measures currently available to the study. Although the OPI level (Dependent Variable) is an ordered-categorical variable, MLR was selected as a statistical model, rather than Ordinal Multiple Regression, because it consists of nine categories and display a normal-shape distribution (*M* = 5.07, *SD* = 2.00, Skewness = -0.02, and Kurtosis = -.74). In this study, predictive power is defined as adjusted R^2, and efficiency is defined as the number of predictors in the model. Model fit to the observed data by each regression model was judged empirically and theoretically. More specifically, the model fit was evaluated first in terms of predictive power and efficiency, and then carefully examined for theoretical plausibility in light of the literature. In other words, theoretical plausibility was not used as an initial evaluation criterion due to its subjectivity. This is what is meant by "empirical-based" in this study. This chapter begins with descriptive statistics of 18 CAF measures (Independent Variables) and the OPI levels (Dependent Variable), followed by underlying data assumptions for MLR. The next section reports on findings from a preliminary analysis with MLR. The fourth section presents on the comparison among different models, and the fifth section reports on the final model. The chapter concludes by discussing the research findings in response to the second research question.

Table 5.1 *Descriptive Statistics for the 18 CAF measures and OPI levels*

	N	Min.	Max.	*M*	*SD*	Variance	Skewness
OPI Levels	170	1.00	9.00	5.07	2.00	3.99	-0.02
9. Phonation time ratio	170	27.65	83.86	61.22	12.38	153.28	-0.40
10. Speech rate	170	88.31	437.32	228.94	71.57	5121.72	0.41
11. Articulation rate	170	171.92	515.39	321.68	69.16	4782.45	0.25
12. Mean length run	170	3.33	23.80	8.88	3.67	13.49	1.27
13. Effective speech rate	170	53.33	419.64	207.80	75.73	5735.72	0.40
14. Effective articulation rate	170	239.36	574.99	366.86	64.74	4191.11	0.60
19. Silent pause ratio	170	8.77	60.81	29.39	11.08	122.70	0.42
20. Silent&Filled pause ratio	170	14.54	72.35	38.37	12.24	149.84	0.41
29. Silent pause ratio within AS	170	1.07	53.74	21.38	10.11	102.12	0.91
30. Silent&Filled pause ratio within AS	170	1.07	72.35	27.45	11.68	136.43	0.95
31. Ratio of silent pause time between AS to total response time	170	0.00	32.77	8.01	5.94	35.31	1.37
32. Ratio of Silent&Filled pause time between AS to total response	170	0.00	42.05	10.92	7.66	58.65	1.32
40. Repeat ratio	170	0.00	11.90	1.54	1.87	3.49	**2.12**
41. Short repetition ratio	170	0.00	6.04	0.84	1.01	1.02	1.96
42. Self-correction ratio	170	0.00	11.80	1.03	1.73	3.00	**3.09**
43. DYSF ratio	170	0.00	17.34	3.41	2.88	8.30	1.57
45. Syntactic complexity	170	0.00	9.00	2.44	1.35	1.83	1.48
48. Error-free AS-unit ratio	170	0.00	100.00	54.44	31.24	975.68	-0.22

Descriptive statistics of the 18 CAF measures and the OPI levels

This section presents descriptive statistics of the variables to be used in MLR analyses. Table 5.1 displays a summary of means, standard deviations (*SD*), observed minimum and maximum values, variances, and skewness for all variables. Overall, most variables are normally distributed. REPEAT RATIO and SELF-CORRECTION RATIO are positively skewed[16]; however, as mentioned in the previous chapter, dysfluency variables tend to be positively skewed due to a small number of occurrences.

Testing of Assumptions for Statistical Analyses

In order to make sure that the data is suitable for MLR analyses, data was evaluated if assumed data conditions were met by submitting all the 18 variables to SPSS. The seven assumptions for MLR have to do with outliers, collinearity of data, independent errors, random normal distribution of errors, homoscedasticity, linearity, and non-zero variances. This section repots on each assumption check.

The first step is checking for outliers. An analysis of standard residuals was carried out. The standard residuals (*Std. Residual Min*.=-2.20, *Std. Residual Max.*= 2.46) showed the values within the 3.29 and -3.29 range; therefore, the data does not contain any outliers.

The next step is checking for collinearity of data. Table 5.2 reports the coefficients table from a SPSS output. A VIF value greater than 10, or a Tolerance value less than 0.1 indicates that the variable is highly correlated with other variables, and it is a multicollinearity issue.

[16] The values are emphasized in bold letters in Table 5.1.

Table 5.2 *SPSS Output of Coefficients Table*

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	-7.386	6.013		-1.228	0.221		
	9. Phonation time ratio	0.072	0.066	0.446	1.097	0.274	**0.010**	**100.538**
	10. Speech rate	-0.003	0.024	-0.119	-0.138	0.891	**0.002**	**451.623**
	11. Articulation rate	0.002	0.015	0.067	0.128	0.898	**0.006**	**166.716**
	12. Mean length run	-0.035	0.046	-0.065	-0.767	0.444	0.228	4.384
	13. Effective speech rate	0.011	0.020	0.406	0.528	0.598	**0.003**	**358.831**
	14. Effective articulation rate	0.007	0.011	0.216	0.635	0.526	**0.014**	**70.552**
	29. Silent pause ratio within AS	-0.084	0.072	-0.424	-1.164	0.246	**0.012**	**80.641**
	30. Silent&Filled pause ratio within AS	0.132	0.088	0.770	1.498	0.136	**0.006**	**160.933**
	31. Ratio of silent pause time between AS to total response time	-0.039	0.075	-0.115	-0.516	0.606	**0.033**	**30.165**
	32. Ratio of Silent&Filled pause time between AS to total response	0.091	0.088	0.350	1.041	0.300	**0.015**	**68.651**
	40. Repeat ratio	-0.194	0.133	-0.182	-1.465	0.145	0.107	9.345
	41. Short repetition ratio	-0.041	0.150	-0.021	-0.272	0.786	0.286	3.501
	42. Self-correction ratio	0.078	0.126	0.068	0.616	0.539	0.137	7.309
	45. Syntactic complexity	0.348	0.073	0.235	4.747	<0.001	0.670	1.492
	48. Error-free AS-unit ratio	0.013	0.003	0.205	4.158	<0.001	0.677	1.478

As emphasized in bold in Table 5.2, some variables demonstrated multicollinearity; however, this was expected. In the previous chapter, the results of Spearman's rank-order correlation among 18

CAF measures revealed that some variables are very strongly correlated to one another. Figure 4.1 in Chapter 4 summarized the relationships among the variables. The variables showing multicollinearity issues displayed in Table 5.2 align with these findings. Among the variables that were found to be strongly correlated in Chapter 4, 19. SILENT PAUSE RATIO and 20. SILENT & FILLED PAUSE RATIO are not included in Table 5.2 because they were not included in the regression model. When variables are found to be insignificantly contributing to predicting the outcome (i.e., the OPI levels in this case), those variables are automatically excluded with the stepwise procedure. Since the variable 9. PHONATION TIME is in the model, which was found to be very strongly correlated with 19. SILENT PAUSE RATIO and 20. SILENT & FILLED PAUSE RATIO, variables 19 and 20 did not contribute enough to stay in the model to add any useful information; hence they were removed from the model. Although this data violates multicollinearity, the purpose of this study is to investigate which variables can be utilized as predictors. Since this study is exploratory in nature, it is more meaningful to keep all the variables first, and then keep adjusting them throughout the process of identifying the final model. Once the final model is determined, this assumption testing will be addressed again to make sure there is no multicollinearity.

In order to test for independent errors, Durbin-Watson value was calculated. The value *(Durbin-Watson value*= 1.63) supports that the data met the assumption. Next, random normally distributed errors, homoscedasticity, and linearity were checked. The histogram of standardized residuals indicated that the data contained normally distributed errors (see Figure 5.1), as did the normal P-P plot of standardized residuals, which showed points that were mostly on the line and some that twined around the line (see Figure 5.2). Also, the scatterplot of standardized predicted values showed that the data met the assumptions of homogeneity of variance and linearity (see Figure 5.3).

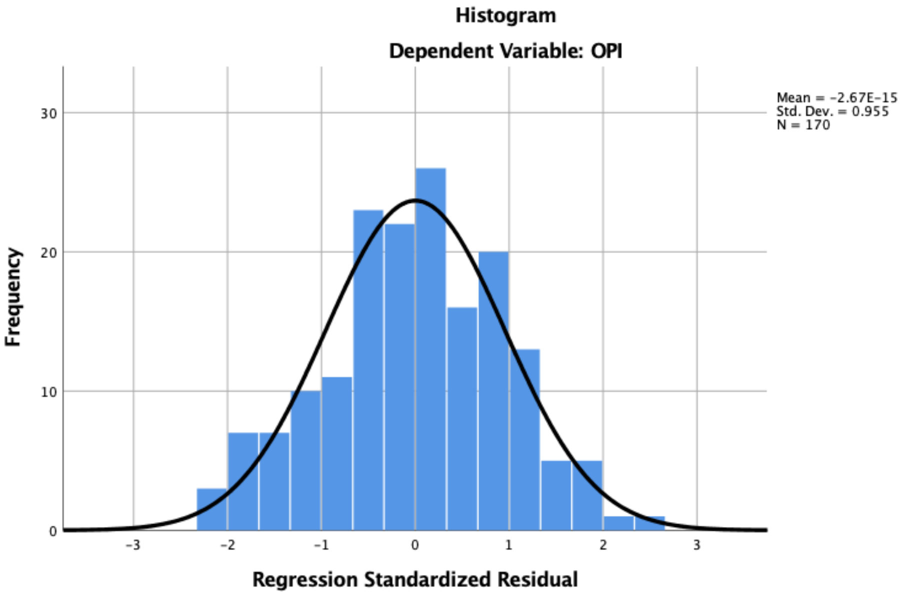

Figure 5.1 *Histogram of Standardized Residuals*

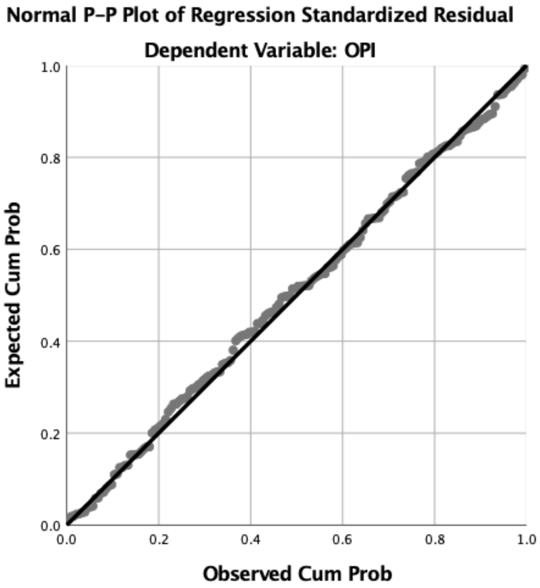

Figure 5.2 *P-P plot of Standardized Residuals*

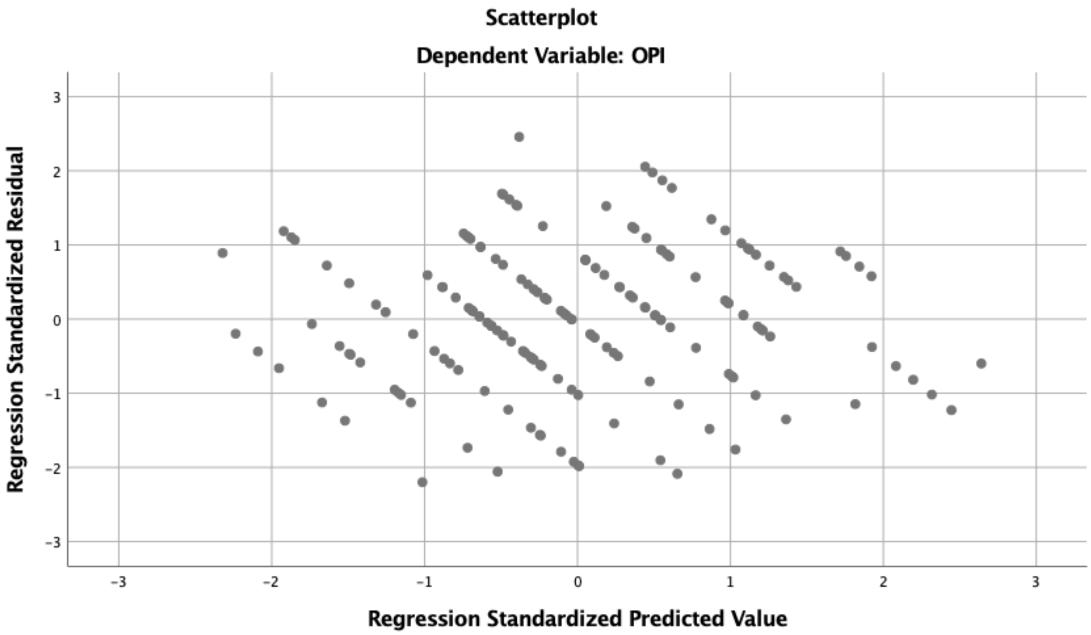

Figure 5.3 *Scatterplot of Standardized Predicted Values*

Lastly, in order to test for non-zero variances, the variance for each variable was computed and presented in Table 5.1. As all the variances are greater than 0, the assumption of non-zero variances is met.

These results support that the data met the assumptions for conducting data analysis with MLR, except for multicollinearity. This problem will be continuously addressed throughout the process of finding the last model.

Results and Discussion

This section reports on the data analysis processes of finding a model that can best predict the OPI levels in the most efficient way, given the current dataset. As mentioned at the beginning of this chapter, this study defines predictive power as adjusted R^2, and efficiency as the least number of predictors. This section carefully explains and justifies the steps taken to reach the final model, and then discusses an interpretation of the results with the final model.

Step 1: Finding the Primary Predictor

As mentioned previously, this study is not testing hypotheses based on specific theories or literature, but rather exploratory. Since its focus is to find a set of good predictors solely from the empirical data, a multiple linear regression using the stepwise approach was selected. This is a preliminary step for finding the primary predictor that contributes the most to the prediction of an outcome, among the 18 CAF variables. In stepwise regression, predictor variables are entered into the regression equation one at a time, based on the predictive power (i.e., adjusted R^2). Therefore, the first predictor that goes into the model has the greatest predictive power.

Results

A stepwise multiple regression was used to evaluate whether all 18 CAF variables were necessary to predict the OPI levels, and if not, which variable has the most predictive power. Table 5.3 shows a summary of the results. At step 1 of the analysis, 13. EFFECTIVE SPEECH RATE entered into the regression equation and was significantly related to the OPI levels F (1, 168) = 269.23, p <.01. The adjusted R^2 was .613, indicating approximately 61.3% of the variance of the OPI levels is accounted for by EFFECTIVE SPEECH RATE. There were five other variables that were added to the model after EFFECTIVE SPEECH RATE; however, since the motivation for the multiple regression analysis was to determine which predictor enters the model first, this section only focuses on the first predictor. Also, it has to be noted that SPSS actually conducted another step after adding the sixth predictor. Although the last step appeared to be free of multicollinearity issues described in the previous section, since the output was computed automatically by SPSS, the processes and rationale behind it are unknowable (see Appendix II). For this reason, the current study considers next to the last step as the starting point of a manual investigation.

Table 5.3 *Results of a Stepwise Multiple Regression*

Variables Entered	Cumulative R^2	R^2 Change	*b*	Beta	*t*	VIF
13. Effective speech rate	**0.61**	0.62	0.001	0.05	0.40	**10.49**
45. Syntactic complexity	0.65	0.04	0.36	0.24	5.22**	1.28
48. Error-free AS-unit ratio	0.68	0.03	0.01	0.20	4.16**	1.36
40. Repeat ratio	0.70	0.02	-0.24	-0.23	-4.66**	1.42
14. Effective articulation rate	0.71	0.02	0.01	0.36	4.19**	4.51
16. Silent pause ratio	0.72	0.01	-0.04	-0.22	-2.90**	3.38
(Constant)			0.66			

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

Discussion

The first variable that was added to the regression equation was 13. EFFECTIVE SPEECH RATE, followed by 45. SYNTACTIC COMPLEXITY, 48. ERROR-FREE AS-UNIT RATIO, and so on. The fact that SPSS stopped adding more variables to the model after a few steps indicates that not all 18 CAF variables were necessary to predict the OPI levels; it is possible to predict the OPI levels with a smaller number of CAF variables.

Although the result of the stepwise multiple regression suggested that EFFECTIVE SPEECH RATE might be the primary predictor, it cannot be concluded yet without further investigation due to noise in the data caused by multicollinearity issues. Given the strong correlation between EFFECTIVE SPEECH RATE and the OPI levels (*r* = .78**, see Table 4.2), it is plausible that EFFECTIVE SPEECH RATE is be the primary predictor; however, the other *Speech Speed* variables are also strongly correlated to the OPI levels, and among themselves. Then, the primary predictor could be any of the *Speech Speed* variables, and therefore needs further investigation.

In Table 5.3, the VIF value emphasized in bold showed that EFFECTIVE SPEECH RATE was collinear with other variables in the model, most likely with EFFECTIVE ARTICULATION RATE. It means that this model is not efficient because the two variables add overlapping information to the model. In order to address this multicollinearity issue and to obtain more accurate output, the highly correlated variables need to be removed.

The result of this analysis provided a very important piece of information that *Speech Speed* related variables might be the key predictor; however, further investigation is necessary to determine which one of the *Speech Speed* variables works the best as a key predictor in a model.

Step 2: Testing for the Primary Predictor

The previous analysis revealed that *Speech Speed* variables contribute the most to predicting the OPI levels. Therefore, as a next step, this section reports on the analyses conducted to compare among the *Speech Speed* variables. For these analyses, a combination of hierarchical and stepwise method was selected. To be more precise, one of the five *Speech Speed* variables was manually entered as the first predictor, and then the rest of 13 CAF measures are entered to the regression model using the stepwise approach. In this way, the multicollinearity issue among the *Speech Speed* variables was resolved. A total of five multiple regressions were conducted, each time with a different *Speech Speed* variable. Then the outcome models were compared in terms of adjusted R^2 and the efficiency of the model.

Results

The first variable to be examined was 10. SPEECH RATE. An analysis with multiple linear regression was conducted to predict the OPI levels with 14 CAF variables. The result of the data analysis is shown in Table 5.4. A significant regression equation was found (F (6, 163) = 75.03, p < .01), with an adjusted R^2 of .724. The predictive model shows that the OPI level is equal to: -1.85 + 0.02 (SPEECH RATE) - 0.25 (REPEAT RATIO) + 0.36 (SYNTACTIC COMPLEXITY) + 0.01 (ERROR-FREE AS-

UNIT RATIO) + 0.08 (SILENT & FILLED PAUSE RATIO) - 0.05 (SILENT PAUSE RATIO). All six variables in the model were significant predictors of the OPI levels. The VIF values of 10. SPEECH RATE, 20. SILENT & FILLED PAUSE RATIO, AND 19. SILENT PAUSE RATIO showed slight multicollinearity; however, the VIF values less than 10 are considered acceptable (Hair et al., 1995).

Table 5.4 *MLR Results for Speech Rate*

Variable	Cumulative R^2	R^2 Change	*b*	Beta	*t*	VIF
10. Speech rate	0.557	0.559	0.02	0.67	8.53**	3.83
40. Repeat ratio	0.633	0.078	-0.25	-0.23	-5.25**	1.16
45. Syntactic Complexity	0.670	0.039	0.36	0.25	5.41**	1.26
48. Error-free AS-unit ratio	0.696	0.028	0.01	0.20	4.45**	1.29
20. Silent & Filled pause ratio	0.708	0.013	0.08	0.47	4.32**	7.29
19. Silent pause ratio	**0.724**	0.017	-0.05	-0.30	-3.26**	5.02
(Constant)			-1.85			

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

Similarly, the other four *Speech Speed* variables were examined following the same procedure used for SPEECH RATE. Significant regression equations were found for the following measures: 11. ARTICULATION RATE (*F* (6, 163) = 74.12, *p* < .01), 12. MEAN LENGTH RUN (*F* (5, 164) = 55.61, *p* < .01), 13. EFFECTIVE SPEECH RATE (*F* (7, 162) = 65.47, *p* < .01), and 14. EFFECTIVE ARTICULATION RATE (*F* (5, 164) = 89.43, *p* < .01). The output models for 11. ARTICULATION RATE, 12. MEAN LENGTH RUN, 13. EFFECTIVE SPEECH RATE, and 14. EFFECTIVE ARTICULATION RATE are displayed in Table 5.5, 5.6, 5.7, and 5.8, respectively.

Table 5.5 *MLR Results for Articulation Rate*

Variable	Cumulative R^2	R^2 Change	*b*	Beta	*t*	VIF
11. Articulation rate	0.435	0.438	0.01	0.49	8.41**	2.07
45. Syntactic Complexity	0.544	0.111	0.38	0.26	5.61**	1.25
40. Repeat ratio	0.604	0.061	-0.24	-0.23	-5.19**	1.16
19. Silent pause ratio	0.660	0.057	-0.12	-0.64	-6.51**	5.88
20. Silent & Filled pause ratio	0.689	0.031	0.08	0.52	4.57**	7.70
48. Error-free AS-unit ratio	**0.722**	0.034	0.01	0.21	4.51**	1.29
(Constant)			-0.59			

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

Table 5.6 *MLR Results for Mean Length Run*

Variable	Cumulative R^2	R^2 Change	*b*	Beta	*t*	VIF
12. Mean length run	0.373	0.337	0.10	0.18	2.44*	2.32
40. Repeat ratio	0.485	0.114	-0.28	-0.26	-5.11**	1.15
45. Syntactic Complexity	0.542	0.060	0.41	0.28	5.17**	1.27
48. Error-free AS-unit ratio	0.594	0.053	0.02	0.26	4.88**	1.27
19. Silent pause ratio	**0.618**	0.025	-0.04	-0.23	-3.34**	2.02
(Constant)			3.93			

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

Table 5.7 *MLR Results for Effective Speech Rate*

Variable	Cumulative R^2	R^2 Change	*b*	Beta	*t*	VIF
13. Effective speech rate	0.613	0.616	0.02	0.75	8.71**	4.64
45. Syntactic complexity	0.652	0.040	0.36	0.24	5.31**	1.29
48. Error-free AS-unit ratio	0.677	0.027	0.01	0.20	4.22**	1.37
40. Repeat ratio	0.695	0.020	-0.13	-0.12	-2.67**	1.33
20. Silent & Filled pause ratio	0.702	0.008	0.08	0.51	4.57**	7.75
19. Silent pause ratio	0.719	0.018	-0.06	-0.31	-3.49**	5.03
42. Self-correction ratio	**0.728**	0.010	0.13	0.11	2.51*	1.28
(Constant)			-2.10			

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

Table 5.8 *MLR Results for Effective Articulation Rate*

Variable	Cumulative R^2	R^2 Change	*b*	Beta	*t*	VIF
14. Effective articulation rate	0.447	0.451	0.01	0.39	8.43**	1.31
19. Silent pause ratio	0.565	0.119	-0.04	-0.24	-5.17**	1.32
40. Repeat ratio	0.654	0.090	-0.25	-0.23	-5.39**	1.15
45. Syntactic complexity	0.691	0.039	0.36	0.24	5.37**	1.25
48. Error-free AS-unit ratio	**0.723**	0.033	0.01	0.20	4.50**	1.23
(Constant)			0.73			

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

Discussion

The results from a series of multiple regression analyses with different *Speech Speed* variables were carefully analyzed by comparing the output models for predictive power and efficiency. Table 5.9 shows a summary of the five models each with a different *Speech Speed* variable. The results show that among them, the model with 14. EFFECTIVE ARTICULATION RATE demonstrated the highest adjusted R^2 value with the least number of predictors. Although the analysis in the previous step suggested that EFFECTIVE SPEECH RATE might be the primary predictor, after controlling for the multicollinearity issue, the results of this analysis revealed that the model with EFFECTIVE ARTICULATION RATE works the best. Considering that EFFECTIVE SPEECH RATE has a stronger correlation with the OPI levels compared with any other *Speed* variables, it would work better if it is used as a solo predictor; however, when the variable is used in combination with other predictors, EFFECTIVE ARTICULATION RATE provides less overlapping information to the model, and therefore, it is more efficient.

Table 5.9 *Summary of the Five MLR Models*

Speed Variable	Model Predictors[17]	No. of Predictors	Adjusted R^2
10. Speech rate	10, 40, 45, 48, 20, 19	6	0.724
11. Articulation rate	11, 45 40, 19, 20, 48	6	0.722
12. Mean length run	12, 40, 45, 48, 19	5	0.618
13. Effective speech rate	13, 45, 48, 40, 20, 19, 42	7	0.728
14. Effective articulation rate	**14, 19, 40, 45, 48**	**5**	**0.723**

The results of the analyses showed that among the five *Speech Speed* variables, EFFECTIVE ARTICULATION RATE works the best as the primary predictor when used in combination with 19. SILENT PAUSE RATIO, 40. REPEAT RATIO, 45. SYNTACTIC COMPLEXITY, and 48. ERROR-FREE AS-UNIT RATIO. Although this output model seems promising, there is one problem; the predictor includes the variable 19. SILENT PAUSE RATIO. Again, as seen in Figure 4.1, this variable is known to be very strongly correlated to *Amount of Speech* variables: 9. PHONATION TIME RATIO and 20. SILENT & FILLED PAUSE TIME RATIO. Although no multicollinearity issues were found in this

[17] The numbers represent the CAF variables names. Please refer to Table 5.1 for the corresponding variable names.

model, it is worthwhile to further investigate which one of the *Amount of Speech* variables works the best in predicting the OPI levels to optimize the model.

Step 3: Adjusting the Model

The results of the previous analyses revealed that EFFECTIVE ARTICULATION RATE is the primary predictor that represents *Speech Speed*; however, the model with this predictor included a variable that very highly correlates with other variables. This section reports on the analyses conducted to compare among the three variables that represent *Amount of Speech*. Similarly to the previous analyses, a total of three multiple regressions with a combination of hierarchical and stepwise approaches were conducted. This time, EFFECTIVE ARTICULATION RATE was manually entered into the model as the first predictor, and then one of the three *Amount of Speech* variables and the rest of 10 CAF variables were submitted to SPSS using the stepwise approach. Then, the outcome models were compared in terms of adjusted R^2 and the number of predictors in the model.

Results

First, 9. PHONATION TIME RATIO was examined. A multiple linear regression was conducted to predict the OPI levels based on 12 CAF variables. The output model is shown in Table 5.10. A significant regression equation was found (F (5, 164) = 83.24, p < .01), with adjusted R^2 of .709. The predictive model shows that the OPI level equal to -2.68 + 0.12 (EFFECTIVE ARTICULATION RATE) + 0.03 (PHONATION TIME RATIO) - 0.23 (REPEAT RATIO) + 0.37 (SYNTACTIC COMPLEXITY) + 0.01 (ERROR-FREE AS-UNIT RATIO). All five variables in the model were significant predictors of the OPI levels.

Table 5.10 *MLR Results for Phonation Time Ratio*

Variable	Cumulative R^2	R^2 Change	*b*	Beta	*t*	VIF
14. Effective articulation rate	0.447	0.451	0.12	0.40	8.37**	1.31
9. Phonation time ratio	0.563	0.117	0.03	0.21	4.13**	1.44
40. Repeat ratio	0.642	0.080	-0.23	-0.23	-5.09**	1.15
45. Syntactic complexity	0.682	0.041	0.37	0.25	5.42**	1.26
48. Error-free AS-unit ratio	**0.709**	0.028	0.01	0.19	4.05**	1.29
(Constant)			-2.68			

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

Second, 19. SILENT PAUSE RATIO was examined. Although this time the highly correlated variables were removed from the data before running the multiple regression, the outcome model was exactly the same as the previous analysis, as shown in Table 5.8.

Lastly, 20. SILENT & FILLED PAUSE TIME RATIO was examined. A multiple linear regression was conducted to predict the OPI levels based on 12 CAF variables. The output model is shown in Table 5.11. A significant regression equation was found (F (6, 163) = 74.10, $p < .01$), with and adjusted R^2 of .722. The predictive model shows that the OPI level equal to 0.73 + 0.01 (EFFECTIVE ARTICULATION RATE) + 0.36 (SYNTACTIC COMPLEXITY) + 0.01 (ERROR-FREE AS-UNIT RATIO) - 0.25 (REPEAT RATIO) - 0.04 (SILENT PAUSE RATIO WITHIN AS-UNIT) - 0.05 (RATIO OF SILENT PAUSE TIME BETWEEN AS-UNIT TO TOTAL RESPONSE TIME). All six variables in the model were significant predictors of the OPI levels.

Table 5.11 *MLR Results for Silent & Filled Pause Ratio*

Variable	Cumulative R^2	R^2 Change	*b*	Beta	*t*	VIF
14. Effective articulation rate	0.447	0.451	0.01	0.39	*8.40***	1.31
45. Syntactic complexity	0.561	0.115	0.36	0.24	5.11**	1.34
48. Error-free AS-unit ratio	0.634	0.075	0.01	0.21	4.37**	1.34
40. Repeat ratio	0.680	0.047	-0.25	-0.23	-5.37**	1.15
29. Silent pause ratio within AS-unit	0.708	0.029	-0.04	-0.22	-4.65**	1.29
31. Ratio of silent pause time between AS-unit to total response time	**0.722**	0.015	-0.05	-0.14	-3.02**	1.26
(Constant)			0.73			

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

Discussion

After conducting three multiple regressions, the output models were compared for predictive power and efficiency. Table 5.12 shows a summary of the three models each with a different *Amount of Speech* variable. The results show that the model with 19. SILENT PAUSE RATIO demonstrated the greatest adjusted R^2 value with the least number of predictors.

Table 5.12 *Summary of the Three MLR Models*

Controlled Variable	Predictors	No. of Predictors	Adjusted R^2
9. Phonation time ratio	14, 9, 40, 45, 48	5	.709
19. Silent pause ratio	**14, 19, 40, 45, 48**	**5**	**.723**
20. Silent & Filled pause ratio	14, 45, 48, 40, 29, 31	6	.722

From these results, this section concludes that the model with the combination of five predictors (i.e., 14. EFFECTIVE ARTICULATION RATE, 19. SILENT PAUSE RATIO, 40. REPEAT RATIO, 45. SYNTACTIC COMPLEXITY, and 48. ERROR-FREE AS-UNIT RATIO) as the final model, and this model can best predict the OPI levels than any other models examined. The next section reports on the final model in detail and discusses interpretations and the validity of the model.

Step 4: The Final Model

This section first reports on the final model once again this time in detail, and then discusses the meaning of the model in relation to the OPI levels.

Explanation of The Model

A series of multiple regressions was conducted to examine which combination of CAF variables can best predict the OPI levels. After comparing several models and controlling for the multicollinearity issues, it was found that a combination of five CAF variables (i.e., 14. EFFECTIVE

ARTICULATION RATE, 19. SILENT PAUSE RATIO, 40. REPEAT RATIO, 45. SYNTACTIC COMPLEXITY, and 48. ERROR-FREE AS-UNIT RATIO) can best predict the OPI levels (F (5, 164) = 89.43, $p < .01$, R^2 = .86, $R^2_{Adjusted}$ = .722). The regression equation is:

The OPI levels = 0.73 + 0.01 (EFFECTIVE ARTICULATION RATE) - 0.04 (SILENT PAUSE RATIO) - 0.25 (REPEAT RATIO) + 0.36 (SYNTACTIC COMPLEXITY) + 0.01 (ERROR-FREE AS-UNIT RATIO).

In this model, the adjusted R^2 value was .723, indicating that approximately 72.3% of the variance of the OPI levels can be explained by these five predictors.

Table 5.13 *Summary of The Final Model*

Variable	Cumulative R^2	R^2 Change	*b*	Beta	*t*	VIF
14. Effective articulation rate	0.447	0.451	0.01	0.39	8.43**	1.31
19. Silent pause ratio	0.565	0.119	-0.04	-0.24	-5.17**	1.32
40. Repeat ratio	0.654	0.090	-0.25	-0.23	-5.39**	1.15
45. Syntactic complexity	0.691	0.039	0.36	0.24	5.37**	1.25
48. Error-free AS-unit ratio	**0.723**	0.033	0.01	0.20	4.50**	1.23
(Constant)			0.73			

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

Table 5.13 shows a summary of the final model. The t-values show that all five variables are significant predictors at $p < .01$ level. The VIF values were all found to be below 3, indicating that the correlations among the predictors are very low; therefore, there is no multicollinearity issue, and each variable makes unique contribution to the model. The standardized coefficient Beta values show that EFFECTIVE ARTICULATION RATE contributes the most to predicting the OPI levels. The three predictors, SILENT PAUSE RATIO, REPEAT RATIO, and SYNTACTIC COMPLEXITY contribute about the same amount, and ERROR-FREE AS-UNIT RATIO makes the least contribution of the five.

Interpretation of The Model

This section examines interpretations of the final model regression equation. The equation shows that as the OPI rating advances, the values of EFFECTIVE ARTICULATION RATE, SYNTACTIC COMPLEXITY, and ERROR-FREE AS-UNIT RATIO increase, but the values in SILENT PAUSE RATIO and REPEAT RATIO decrease. This means that as the OPI rating advances, examinees can produce more meaningful, complex, and accurate speech at a faster rate, with less planning time (Skehan and Foster, 1999) and repetition.

In the final model, each predictor makes unique contribution to predicting the outcome, representing different aspects of speech. Table 5.14 shows a list of CAF measures used for the analyses and their categorization. EFFECTIVE ARTICULATION RATE represents *Speech Speed* and *Density*, SYNTACTIC COMPLEXITY represents *Complexity* of speech, and ERROR-FREE AS-UNIT represents *Accuracy* of speech in terms of grammar and vocabulary. Similarly, SILENT PAUSE RATIO can be seen as representing both *Amount of Speech* and speech planning time, and REPETITION RATIO represents *Dysfluency*. As a composite predictor, this model captures oral proficiency in a multidimensional way, each predictor representing an important aspect of speech production.

Table 5.14 *Aspects of Speech Category and CAF Variables*

Category	Sub-category	Measures
Speed		10. Speech rate
		11. Articulation rate
	Speech density	12. Mean length run
		13. Effective speech rate
		14. Effective articulation rate[18]

[18] Bolded variables are the predictors in the final model.

Table 5.14 continued

Pause	*Amount of Speech*	9. Phonation time ratio
		19. Silent pause ratio
		20. Silent and filled pause ratio
	Pause Location	29. Silent pause ratio within AS
		30. Silent and filled pause ratio within AS
		31. Ratio of silent pause time between AS to total response time
		32. Ratio of silent and filled pause time between AS to total response time
Dysfluency		**40. Repeat ratio**
		41. Short repetition ratio
		42. Self-correction ratio
		43. DYSF ratio
Complexity		**45. Syntactic complexity**
Accuracy		**48. Error-free AS-unit ratio**

Interestingly, some variables were not selected as predictors. For example, in the *Speed* category in Table 5.14, although all five variables showed strong correlations (above *r* = .65) with the OPI levels (see Table 4.2), 14. EFFECTIVE ARTICULATION RATE was found to be the best predictor of all. In fact, 10. SPEECH RATE and 14. EFFECTIVE SPEECH RATE had actually demonstrated higher correlation coefficients (*r* = .74** and *r* = .78**, respectively) than EFFECTIVE ARTICULATION RATE (*r* = .67**) in the analysis of Chapter 4. The reason why EFFECTIVE ARTICULATION RATE was found to be a better predictor for the final model can be attributed to the meaning of this measure. As explained in Chapter 3, the calculation for this measure is ((effective syllable count) / (Speech time – Dysfluency time)) *60, which represents how fast a speaker can produce effective syllables if s/he is not interrupted by any pauses or dysfluency phenomena. Then, it makes sense that the final model included variables that represent pause (i.e., SILENT PAUSE RATIO) and dysfluency phenomenon (i.e., REPEAT RATIO), because EFFECTIVE ARTICULATION RATE only captures the best performance of a speaker. On the other hand, all other *Speed* variables include some portion of pause or dysfluency phenomena in their calculation, overlapping with other variables that represent pause or dysfluency. When EFFECTIVE ARTICULATION RATE is used as a predictor in combination with SILENT PAUSE RATIO and REPEAT RATIO, the correlation coefficient reaches r = .81 (see Appendix III), which is higher than those of SPEECH RATE and EFFECTIVE SPEECH RATE. In the literature, Mean Length Run was found to be a good predictor of L2 oral proficiency; however, the results of this study revealed that this measure showed the least contribution when compared to the other *Speech Speed* variables (see Table 5.9). This measure might be a good predictor as a stand-alone predictor, but when used in combination with other CAF variables, it did not work as well as others.

The three variables in the *Amount of Speech* category are also very strongly correlated with each other (see Figure 4.1). What is intriguing about *Pause* related variables is that none of the *Pause Location* variables were included in the final model. According to the data used in this study, the analyses revealed that although the amount of pause (i.e., SILENT PAUSE RATIO) matters, the location of pause does not matter as much in predicting the OPI levels.

For *Dysfluency* variables, it is interesting to see that REPEAT RATIO was selected among the four. Table 4.2 in Chapter 4 shows that while SHORT REPETITION RATIO and SELF-CORRECTION RATIO showed non-significant correlations with the OPI levels, REPEAT RATIO showed moderate correlation (r = -.41**). This might be the reason why REPEAT RATIO was selected for the model over the others to represent *Dysfluency.*

It is worth pointing out that the *fluency* variables (i.e., EFFECTIVE ARTICULATION RATE, SILENT PAUSE RATIO, and REPETITION RATIO) in the final model align with Skehan's (2009) proposed three categories of fluency: 1) speed, 2) breakdown, and 3) repair. EFFECTIVE

ARTICULATION RATE represents speed (and density), SILENT PAUSE RATIO for breakdown, and REPETITION RATIO for repair. In this sense, it can be concluded that these three variables represent *fluency* well.

For *Complexity* and *Accuracy*, there was only one variable for each category, but both variables, 45. SYNTACTIC COMPLEXITY and 48. ERROR-FREE AS-UNIT RATIO, were included as significant predictors in the final model. Also, all the multiple regression models that were shown in this chapter always included both variables, suggesting that they are very important predictors of the OPI levels. As in the literature (e.g. Housen, Kuiken & Vedder, 2012; Larsen-Freeman, 1978; Skehan 1996), the findings of this study support that *Complexity*, *Accuracy*, and *Fluency* are indeed important components of oral proficiency, and each makes unique contribution to predicting L2 oral proficiency, to the extent that the OPI levels accurately represent it.

This section first reported on the final model and then discussed its interpretations in terms of the OPI levels. After a thorough discussion of the final model, it is safe to conclude that this model captures oral proficiency in a meaningful and multidimensional way.

Response to Research Question 2

This section answers Research Question 2 “Which combination of CAF variables can best predict examinees’ L2 oral proficiency levels?” A series of multiple regressions was conducted and compared. As the final model, it was found that the following five CAF variables can best predict examinees’ L2 oral proficiency levels: EFFECTIVE ARTICULATION RATE, SILENT PAUSE RATIO, REPEAT RATIO, SYNTACTIC COMPLEXITY, and ERROR-FREE AS-UNIT RATIO. The relationship among the variables is explained by the following multiple regression equation:

The OPI levels = 0.73 + 0.01 (Effective Articulation rate) - 0.04 (Silent pause ratio) - 0.25 (Repeat ratio) + 0.36 (syntactic complexity) + 0.01 (Error-free AS-unit ratio).

Approximately 72.3% of the variance of the OPI levels can be explained by these five predictors.

CHAPTER 6: CONCLUSION

Summary of Findings

This study is one of the first to investigate the possibility of using CAF measures as a composite predictor of oral proficiency. As a first step, Research Question 1 "Which CAF variables correlate with L2 Japanese proficiency levels measured by the ACTFL OPI, and to what extent do they correlate?" was examined. Spearman's rank-order correlation coefficients were computed to investigate the relationship between the 18 CAF measures and the ACTFL OPI levels, as well as among the CAF measures. Results revealed that all *Speed*-related variables and a *Complexity* variable were strongly correlated to the OPI levels. Among all, EFFECTIVE SPEECH RATE, which represents *Speech Speed/Density,* showed the highest correlation coefficient with the OPI levels (*r* = .78). Moderately strong correlations were found between the following measures and the OPI levels: *Speech Quantity, Pause*, *Pause Location* (i.e., SILENT PAUSE RATIO WITHIN AS-UNIT), *Dysfluency* (i.e., REPEAT RATIO), and *Accuracy.* The results indicate that as examinees' oral proficiency advances, they can speak faster, can speak more with meaningful syllables, as well as more complex and accurate sentences that do not contain frequent repetitions or pauses (especially silent pauses within AS-unit). Moreover, among the 18 CAF measures, it was found that the measures in the *Speech Speed* group (i.e., SPEECH RATE, ARTICULATION RATE, MEAN LENGTH OF RUN, EFFECTIVE SPEECH RATE, and EFFECTIVE ARTICULATION RATE), the *Amount of Speech* group (i.e., PHONATION TIME RATIO, SILENT PAUSE RATIO, and SILENT & FILLED PAUSE RATIO), and the *Pause Location* group (i.e., PAUSE LOCATION WITHIN AS-UNIT, and PAUSE BETWEEN AS-UNIT) are very strongly correlated with one within each group.

In order to investigate Research Question 2 "Which combination of CAF variables can best predict examinees' L2 proficiency levels?," a series of multiple linear regression analyses was conducted. Results revealed that a combination of the following five CAF variables can best predict examinees' L2 oral proficiency levels: EFFECTIVE ARTICULATION RATE, SILENT PAUSE RATIO, REPEAT RATIO, SYNTACTIC COMPLEXITY, and ERROR-FREE AS-UNIT RATIO. The following multiple regression equation explains the relationship between the group of variables and the OPI levels:

The OPI levels = 0.73 + 0.01 (EFFECTIVE ARTICULATION RATE) - 0.04 (SILENT PAUSE RATIO) - 0.25 (REPEAT RATIO) + 0.36 (SYNTACTIC COMPLEXITY) + 0.01 (ERROR-FREE AS-UNIT RATIO).

This study found that these five CAF measures can explain approximately 72.3% of the variance of the OPI levels. It is worth pointing out that the final regression model ended up including variables that correspond to Skehan's (2009) proposed three categories of fluency (speed, breakdown, and repair) and variables that represent complexity, accuracy, and fluency, supporting the literature (e.g., Larsen-Freeman, 1978, Skehan, 1996).

Limitations and Future Research

One limitation of this study is the small number of measures used for representing *Complexity* and *Accuracy.* Since the main focus of this study was on fluency-related variables, only one variable each was selected to represent *Complexity* and *Accuracy*. The results of the current study provide encouraging information that CAF are indeed important components of oral proficiency, and all three of them must be taken into account when predicting L2 oral proficiency. Since the simple measures used in this study can make such contributions to the composite model, more fine-grained and sophisticated variables representing *Complexity* and *Accuracy* might be able to yield a multiple regression model with even greater predictive power. Further investigation in this direction may be fruitful.

Similar to the previous point, perhaps including variables other than CAF, such as appropriateness, coherence, elegance of speech, might provide some useful information to the current regression model. For example, when coding, the current researcher noticed that higher OPI level examinees can adjust their speech styles according to who they are talking to (e.g., when talking to themselves to think aloud and when they are speaking to the tester).

Another potential limitation of the current study is the length of speech samples. The speech samples used for this study was an approximately one-minute long excerpt from a 30-minute interview session. Since such a short speech sample length could yield a regression model that can explain 72.3% of the OPI rating variability, it is possible that longer speech sample length yields greater predictive power. Further investigation on different speech sample lengths will be beneficial as well.

Implications

The primary aim of this study was to contribute to the fields of language assessment, oral proficiency, and fluency. The current study provides strong support for the notion that *Complexity, Accuracy*, and *Fluency* must be considered together, not only in L2 pedagogy and research (Larsen-Freeman, 2009), but also in L2 assessment. The findings from this study shed light on future development of an (semi-)automated scoring/grading system for speaking tests. Although the composite predictor model found in this study still needs further refinement, it can explain 72.3% of the OPI levels with five CAF measures based on a speech sample of approximately one minute. It is obvious that there must be follow-up research to refine the model to obtain higher predictive power, but once a model with sufficient predictive power is found, it can offer a great benefit to classrooms, institutions, or high-stake tests, because it can reduce time and cost for conducting and grading/rating speaking tests. One of the greatest burdens of conducting speaking tests is that raters must be trained in order to provide reliable scores. Right now, the most common way is to use two or more human raters, but if an automated scoring system is used to support human raters, it can reduce the cost of human labor, time, and money.

REFERENCES

ACTFL. (2012). ACTFL Proficiency Guidelines. Hastings-on-Hudson, NY: *American Council on the Teaching of Foreign Languages*.

Boersma, P., & Weenink, D. (2017). Praat: doing phonetics by computer [Computer program]. Version 6.0.35, retrieved 16 October 2017 from http://www.praat.org/

Chambers, F. (1997). What do we mean by fluency?. *System*, 25(4), 535-544.

Quené, H., Persoon, I., & De Jong, N. (2010) Syllable Nuclei v2, retrieved 22 February 2019 from https://sites.google.com/site/speechrate/Home/praat-script-syllable-nuclei-v2.

Dandonoli, P., & Henning, G. (1990). An investigation of the construct validity of the ACTFL proficiency guidelines and oral interview procedure. *Foreign Language Annals*, 23(1), 11-22.

Enright, M. K., & Quinlan, T. (2010). Complementing human judgment of essays written by English language learners with e-rater® scoring. *Language Testing*, 27(3), 317-334.

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

Freed, B. (2000). Is fluency, like beauty, in the eyes (and ears) of the beholder?. In *Perspectives on fluency* (pp. 243-265). University of Michigan.

Fukada, A., Hirotani, & M., Matsumoto, K. (2019). Fluency Calculator (Version 1.0.2) A Tool for Calculating Fluency–related Measures [Software], retrieved 27 February 2018 from http://tell.cla.purdue.edu/fluency-calculator/

Fulcher, G. (2003). Testing second language speaking. *Pearson Education*.

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.

Greene, J. O., & Cappella, J. N. (1986). Cognition and talk: The relationship of semantic units to temporal patterns of fluency in spontaneous speech. *Language and Speech*, 29(2), 141-157.

Hirotani, M., Matsumoto, K., & Fukada, A. (2017). The Validity of General L2 Proficiency Tests as Oral Proficiency Measures: A Japanese Learner Corpus Based Study. *Japanese Language & Literature*, 51(2), 243 – 270.

Hirotani, M. (2009). Synchronous versus asynchronous CMC and transfer to Japanese oral performance. *Calico Journal*, 26(2), 413 - 438.

Hirotani, M., Cantrell, K. M., & Fukada, A. (2017). Examination of the validity of J-CAT and SPOT as a placement exam. *Proceedings of the 23rd Princeton Japanese Pedagogy Forum*. 346 – 358.

Housen, A., Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics*, 21, 354-375.

Housen, A., Kuiken, F., & Vedder, I. (Eds.). (2012). Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA (Vol. 32). *John Benjamins Publishing*.

Houston, S. M. (2016). A longitudinal study of the development of fluency of novice Japanese learners: Analysis using objective measures. (Master's thesis). Retrieved from *Purdue e-pub*. https://docs.lib.purdue.edu/open_access_theses/776

Hunt, K. W. (1965). Grammatical structures written at three grade levels (No. 3). Champaign, IL: National Council of Teachers of English.

Ishizaki, A. (2004). What Effect Do Pause 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., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct?. *Applied linguistics*, 29(1), 24-49.

Koizumi, R. (2005). Speaking performance measures of fluency, accuracy, syntactic complexity, and lexical complexity. *JABAET Journal*, 9, 5-34.

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

Lado, R. (1961). Language Testing; the Construction and Use of Foreign Language Test. *Longmans*.

Larsen-Freeman, D. (1978). An ESL index of development. *TESOL quarterly*, 439-448.

Larsen-Freeman, D. (2009). Adjusting expectations: The study of complexity, accuracy, and fluency in second language acquisition. *Applied linguistics*, 30(4), 579-589.

Leclercq, P., Edmonds, A., & Hilton, H. (Eds.). (2014). Measuring L2 proficiency: Perspectives from SLA (Vol. 78). *Multilingual Matters*.

Lennon, P. (1990). Investigating fluency in EFL: a quantitative approach. *Language Learning* 40, 387~,17.

Magnan, S. S. (1986). Assessing speaking proficiency in the undergraduate curriculum: Data from French. *Foreign Language Annals*, 19(5), 429-438.

Mehnert, U. (1998). The effects of different lengths of time for planning on second language performance. *Studies in second language acquisition*, 20(1), 83-108.

Mohle, D. (1984). A comparison of the second language speech production of different native speakers. In HW Dechert, D Mohle, and M Raupach (eds.), *Second language production* (pp. 26–49). Tubingen, German: Guner Narr Verlag.

The National Institute for Japanese Language and Linguistics. (2009). L2 Japanese learners' conversation database [Online database]. retrieved 27 October 2018 from https://nknet.ninjal.ac.jp/nknet/ndata/opi/

Ortega, L. (1999). Planning and focus on form in L2 oral performance. *Studies in second language acquisition*, 21(1), 109-148.

Park, S. (2016). Measuring fluency: Temporal variables and pausing patterns in L2 English speech. (Doctoral dissertation). Retrieved from https://docs.lib.purdue.edu/open_access_dissertations/692

Pallotti, G. (2009). CAF: Defininf, refining and differentiating constructs. *Applied Linguistics*, 30, 590-601.

Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. *Language and communication*, 191, 225.

Préfontaine, Y., & Kormos, J. (2015). The relationship between task difficulty and second language fluency in French: A mixed methods approach. *The Modern Language Journal*, 99(1), 96-112.

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

Sakuragi, T. (2011). The construct validity of the measures of complexity, accuracy, and fluency: Analyzing the speaking performance of learners of Japanese. *JALT Journal*, 33(2), 157-174.

Sajavaara, K., & Lehtonen, J. (1978). Spoken Language and the Concept of Fluency. *Language Centre News*, 1, 23-57.

Schmidt, R. (1992). Psychological mechanisms underlying second language fluency. SSLA 14, 357-385.

Segalowitz, N. (2010). Cognitive bases of second language fluency. *Routledge*.

Skehan, P. (1989). Language testing part II. *Language Teaching*, 22(1), 1-13.

Skehan, P. (1996). Second language acquisition and task-based instruction. In J. Willis, & D. Willis (Eds.) *Challenge and change in language teaching* (pp.17 – 30). Oxford: Heinemann.

Skehan, P., & Foster, P. (1997). The influence of planning and post-task activities on accuracy and complexity in task-based learning. *Language Teaching Research*, 1(3), 185-211.

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

Skehan, P., & Foster, P. (1999). The influence of task structure and processing conditions on narrative retellings. *Language learning*, 49(1), 93-120.

Skehan, P. (2003) Task-based instruction. *Language Teaching* 36, 1–14.

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied linguistics*, 30(4), 510-532.

Surface, E. A., & Dierdorff, E. C. (2003). Reliability and the ACTFL Oral Proficiency Interview: Reporting indices of interrater consistency and agreement for 19 languages. *Foreign Language Annals*, 36(4), 507-519.

Tamaoka, K., & Terao, Y. (2004). Mora or syllable? Which unit do Japanese use in naming visually presented stimuli?. *Applied Psycholinguistics*, 25(1), 1-27.

Thompson, I. (1995). A study of interrater reliability of the ACTFL oral proficiency interview in five European languages: Data from ESL, French, German, Russian, and Spanish. *Foreign Language Annals,* 28(3), 407-422.

Thompson, G. L., Cox, T. L., & Knapp, N. (2016). Comparing the OPI and the OPIc: The effect of test method on oral proficiency scores and student preference. *Foreign Language Annals*, 49(1), 75-92.

Towell, R., Hawkins, R., & Bazergui, N. (1996). The development of fluency in advanced learners of French. *Applied linguistics*, 17(1), 84-119.

Watanabe, S. (1998). Concurrent validity and application of the ACTFL Oral Proficiency Interview in a Japanese language program. *The Journal of the Association of Teachers of Japanese*, 32(1), 22-38.

Xi, X. (2010). Automated scoring and feedback systems: Where are we and where are we heading? *Language Testing*. 27 (3) 291 – 300.

APPENDIX I. A SAMPLE SPEECH TRANSCRIPTION

Sample speech of an examinee (OPI level = Intermediate-High level)

“T” indicates a tester, and “I” indicates an examinee.

Ｔ：えー，そうねじゃあ，韓国の家と，日本の家と，比べると，どんなところが違うと思いますか

Ｉ：今のアパートは〈ん〉，韓国のアパット［アパート］と比べて〈ん〉，うるさいですね〈んんん〉，隣の家の人の〈ん〉はなしー［話］が，ほとんど聞こえます〈ん〉，韓国にいるときは〈ん〉，じぇんじぇん［全然］，経験しないことなので〈ん〉，びっくりしました

Ｔ：あー，そうですか，ん，ほかに，生活スタイルとかはどうですか

APPENDIX II. SPSS OUTPUT

***Model Summary*[h]**

					Change Statistics					
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	Sig. F Change	Durbin-Watson
1	.785[a]	.616	.613	1.243	.616	269.232	1	168	.000	
2	.810[b]	.656	.652	1.180	.040	19.427	1	167	.000	
3	.826[c]	.683	.677	1.136	.027	14.086	1	166	.000	
4	.838[d]	.702	.695	1.103	.020	10.965	1	165	.001	
5	.847[e]	.718	.709	1.077	.016	9.057	1	164	.003	
6	.856[f]	.732	.722	1.054	.014	8.420	1	163	.004	
7	.855[g]	.732	.723	1.051	.000	.163	1	163	.687	1.560

a. Predictors: (Constant), @13.Effective_speech_rate

b. Predictors: (Constant), @13.Effective_speech_rate, @45.syntactic_complexity

c. Predictors: (Constant), @13.Effective_speech_rate, @45.syntactic_complexity, @48.errorfree_ASunit_ratio

d. Predictors: (Constant), @13.Effective_speech_rate, @45.syntactic_complexity, @48.errorfree_ASunit_ratio, @40.repeat_ratio

e. Predictors: (Constant), @13.Effective_speech_rate, @45.syntactic_complexity, @48.errorfree_ASunit_ratio, @40.repeat_ratio, @14.Effective_articulation_rate

f. Predictors: (Constant), @13.Effective_speech_rate, @45.syntactic_complexity, @48.errorfree_ASunit_ratio, @40.repeat_ratio, @14.Effective_articulation_rate, @19.Silent_pause_ratio

g. Predictors: (Constant), @45.syntactic_complexity, @48.errorfree_ASunit_ratio, @40.repeat_ratio, @14.Effective_articulation_rate, @19.Silent_pause_ratio

h. Dependent Variable: OPIS

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	415.735	1	415.735	269.232	.000[b]
	Residual	259.418	168	1.544		
	Total	675.153	169			
2	Regression	442.768	2	221.384	159.095	.000[c]
	Residual	232.385	167	1.392		
	Total	675.153	169			
3	Regression	460.945	3	153.648	119.069	.000[d]
	Residual	214.208	166	1.290		
	Total	675.153	169			
4	Regression	474.293	4	118.573	97.404	.000[e]
	Residual	200.860	165	1.217		
	Total	675.153	169			
5	Regression	484.805	5	96.961	83.540	.000[f]
	Residual	190.348	164	1.161		
	Total	675.153	169			
6	Regression	494.155	6	82.359	74.170	.000[g]
	Residual	180.998	163	1.110		
	Total	675.153	169			
7	Regression	493.974	5	98.795	89.427	.000[h]
	Residual	181.179	164	1.105		
	Total	675.153	169			

a. Dependent Variable: OPIS

b. Predictors: (Constant), @13.Effective_speech_rate

c. Predictors: (Constant), @13.Effective_speech_rate, @45.syntactic_complexity

d. Predictors: (Constant), @13.Effective_speech_rate, @45.syntactic_complexity, @48.errorfree_ASunit_ratio

e. Predictors: (Constant), @13.Effective_speech_rate, @45.syntactic_complexity, @48.errorfree_ASunit_ratio, @40.repeat_ratio

f. Predictors: (Constant), @13.Effective_speech_rate, @45.syntactic_complexity, @48.errorfree_ASunit_ratio, @40.repeat_ratio, @14.Effective_articulation_rate

g. Predictors: (Constant), @13.Effective_speech_rate, @45.syntactic_complexity, @48.errorfree_ASunit_ratio, @40.repeat_ratio, @14.Effective_articulation_rate, @19.Silent_pause_ratio

h. Predictors: (Constant), @45.syntactic_complexity, @48.errorfree_ASunit_ratio, @40.repeat_ratio, @14.Effective_articulation_rate, @19.Silent_pause_ratio

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Correlations			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	.767	.279		2.749	.007					
	@13.Effective_speech_rate	.021	.001	.785	16.408	.000	.785	.785	.785	1.000	1.000
2	(Constant)	.480	.273		1.760	.080					
	@13.Effective_speech_rate	.018	.001	.692	13.816	.000	.785	.730	.627	.822	1.216
	@45.syntactic_complexity	.326	.074	.221	4.408	.000	.512	.323	.200	.822	1.216
3	(Constant)	.270	.269		1.006	.316					
	@13.Effective_speech_rate	.016	.001	.591	10.732	.000	.785	.640	.469	.629	1.589
	@45.syntactic_complexity	.369	.072	.250	5.112	.000	.512	.369	.223	.802	1.247
	@48.errorfree_ASunit_ratio	.012	.003	.188	3.753	.000	.483	.280	.164	.761	1.313
4	(Constant)	.912	.325		2.806	.006					
	@13.Effective_speech_rate	.014	.001	.543	9.791	.000	.785	.606	.416	.586	1.707
	@45.syntactic_complexity	.345	.071	.233	4.889	.000	.512	.356	.208	.793	1.261
	@48.errorfree_ASunit_ratio	.011	.003	.171	3.484	.001	.483	.262	.148	.752	1.329
	@40.repeat_ratio	-.166	.050	-.156	-3.311	.001	-.479	-.250	-.141	.818	1.223
5	(Constant)	-.517	.571		-.905	.367					
	@13.Effective_speech_rate	.009	.002	.351	4.187	.000	.785	.311	.174	.245	4.084
	@45.syntactic_complexity	.368	.069	.249	5.310	.000	.512	.383	.220	.783	1.277
	@48.errorfree_ASunit_ratio	.011	.003	.178	3.714	.000	.483	.279	.154	.751	1.332
	@40.repeat_ratio	-.190	.050	-.178	-3.823	.000	-.479	-.286	-.159	.797	1.254
	@14.Effective_articulation_rate	.007	.002	.215	3.009	.003	.671	.229	.125	.336	2.976

Appendix II continued

6	(Constant)	.661	.690		.957	.340					
	@13.Effective_speech_rate	.001	.003	.053	.404	.687	.785	.032	.016	.095	10.486
	@45.syntactic_complexity	.355	.068	.240	5.220	.000	.512	.378	.212	.780	1.282
	@48.errorfree_ASunit_ratio	.013	.003	.196	4.156	.000	.483	.310	.169	.737	1.357
	@40.repeat_ratio	-.241	.052	-.225	-4.663	.000	-.479	-.343	-.189	.705	1.419
	@14.Effective_articulation_rate	.011	.003	.361	4.191	.000	.671	.312	.170	.222	4.505
	@19.Silent_pause_ratio	-.039	.013	-.216	-2.902	.004	-.567	-.222	-.118	.296	3.375
7	(Constant)	.728	.668		1.089	.278					
	@45.syntactic_complexity	.359	.067	.243	5.365	.000	.512	.386	.217	.800	1.250
	@48.errorfree_ASunit_ratio	.013	.003	.202	4.503	.000	.483	.332	.182	.812	1.232
	@40.repeat_ratio	-.250	.046	-.234	-5.388	.000	-.479	-.388	-.218	.869	1.150
	@14.Effective_articulation_rate	.012	.001	.390	8.425	.000	.671	.550	.341	.763	1.310
	@19.Silent_pause_ratio	-.043	.008	-.240	-5.169	.000	-.567	-.374	-.209	.761	1.315

a. Dependent Variable: OPIS

APPENDIX III. SPSS OUTPUT FOR THE FINAL MODEL

Model Summary[f]

					Change Statistics					
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	Sig. F Change	Durbin-Watson
1	.671[a]	.451	.447	1.486	.451	137.749	1	168	.000	
2	.755[b]	.570	.565	1.319	.119	46.300	1	167	.000	
3	**.812**[c]	.660	.654	1.176	.090	43.907	1	166	.000	
4	.836[d]	.698	.691	1.111	.039	21.168	1	165	.000	
5	.855[e]	.732	.723	1.051	.033	20.275	1	164	.000	1.560

a. Predictors: (Constant), @14.Effective_articulation_rate

b. Predictors: (Constant), @14.Effective_articulation_rate, @19.Silent_pause_ratio

c. Predictors: (Constant), @14.Effective_articulation_rate, @19.Silent_pause_ratio, @40.repeat_ratio

d. Predictors: (Constant), @14.Effective_articulation_rate, @19.Silent_pause_ratio, @40.repeat_ratio, @45.syntactic_complexity

e. Predictors: (Constant), @14.Effective_articulation_rate, @19.Silent_pause_ratio, @40.repeat_ratio, @45.syntactic_complexity, @48.errorfree_ASunit_ratio

f. Dependent Variable: OPIS

VITA

Mayu Miyamoto

EDUCATION

DOCTOR OF PHILOSOPHY, Purdue University**,** *Applied Linguistics*, 2015 – 2019
Dissertation title:
"*Capturing L2 Oral Proficiency with CAF Measures as Predictors of the ACTFL OPI Rating"*
Committee: Atsushi Fukada, Mariko Wei, Jessica Sturm and April Ginther

MASTER OF ARTS, Purdue University, *Japanese Pedagogy/ SLA*, 2011 – 2014
Thesis title: "*Effects of Online Oral Practice on Japanese Pitch Accentuation Acquisition.*"
Committee: Atsushi Fukada, Mariko Wei, and April Ginther
Graduate Certificate in Teaching English as a Second Language, Purdue University 2013 – 2014

BACHELOR OF ARTS, Nanzan University (Japan), *Foreign Studies*, 2007 – 2011
Study Abroad: University of North Carolina at Greensboro (2009)

APPOINTMENTS

Teaching Assistant of Japanese, Purdue University	2017 – present
Placement Test Administration Assistant, Purdue University	2018 summer
Research Assistant for Online Placement Test Development, Purdue University	2018 summer
Research Assistant for Online Performance-Based Test, Purdue University	2015 – 2017
Lecturer of Japanese, Purdue University	2014 – 2015
Teaching Assistant of Japanese, Purdue University	2011 – 2014

RESEARCH INTERESTS

Teaching Japanese as a Second/Foreign Language
Language Testing
Assessment of Oral Proficiency
Second Language Acquisition of Phonetics and Prosody
Computer Assisted Language Learning

PUBLICATIONS

Miyamoto, M., & Fukada, A. (2019). 「継続的オーラルアセスメントの開発–『話せる』を実感する評価法を目指して」. ひつじ書房 (Submitted as a book chapter)

Jessica, S., Suzuki, N., & Miyamoto, M. (2018). French Student and Teacher Attitudes Toward Pronunciation. *The French Review*.

Miyamoto, M., Suzuki, N., Fukada, A., Huang, Y., Hong, S., & Wei, H. (2017). "Teaching Languages Online: Innovations and Challenges" (2017). *Purdue Languages and Cultures Conference 2017.*

Suzuki, N., Okamoto, T., & Miyamoto, M. (2017). 「日本の魅力「再」発見！: リサーチプロジェクトで深める異文化理解」.(Discovering Japan's hidden gems: Enhancing learners' intercultural awareness through a research project). In *Proceedings of the 23rd Princeton Japanese Pedagogy Forum* (pp.131-140).

Miyamoto, M., Suzuki, N., & Fukada, A. (2016).「アニメーションテロップを使用したオンライン復唱練習が発音習得に与える効果：アクセントと特殊拍を中心に」. In N. Takahashi (Ed.), *Proceedings of the 2016 Symposium on Japanese Language Education (AJE)*.

Suzuki, N., & Miyamoto, M. (2016). Effects of online repetition practice with animated visual aid on the acquisition of Japanese pitch accent and special moras. In J. Gao (Ed.), *Proceedings of the 1st Purdue Languages and Cultures Conference*.

Mishima, H., Miyamoto, M., & Yanagisawa, S. (2015). 「スピーキングを重視したオンライン日本語コースの設計・開発・運用」. In S. Sato (Ed.), *Proceedings of The 22nd Princeton Japanese Pedagogy Forum* (pp. 191-205).

Invited Presentations

Miyamoto, M., & Okamoto, T. (2017). Re-evaluating classroom assessments --- What is a good test item? Paper presented at *Washington Associating of Teachers of Japanese.* Seattle, WA.

Conference Presentations

Miyamoto, M., Fukada, A., Sturm, J., & Sundquist, C., (2017). Online Performance-Based Assessments as Routine Achievement Tests. *Paper presented at Annual Convention & World Languages EXPO*. Nashville, TN

Miyamoto, M., & Wei, M., (2017). Effects of Performance-based Achievement Testing on Oral Proficiency. *Paper presented at Annual Convention & World Languages EXPO*. Nashville, TN

Miyamoto, M., & Suzuki, N.,& Fukada, A. (2017). Online oral practice platform Speak Everywhere for daily pronunciation practice. Poster presented at *the Pronunciation in Second Language Learning & Teaching (PSLLT) 9th Annual Conference*. Salt Lake City, UT.

Miyamoto, M., & Fukada, A. (2017).「日本語で何ができるか」を測る—パフォーマンスベースのオンライン到達度テストと採点システム」. Paper presented at *the 7th International Conference on Computer Assisted Systems for Teaching & Learning Japanese.* Tokyo, Japan.

Miyamoto, M. (2016). A Scale Development Project for Performance-Based Test (PBT). Poster presented at *Midwest Association of Language Testers 18th Annual Conference.* West Lafayette, Indiana.

Okamoto, T., & Miyamoto, M. (2016). Developing 21st century skills using Social Networking Services (SNS) as a learning tool. Paper presented at *WAFLT-COFLT Fall 2016 conference.* Portland, Oregon.

Miyamoto, M., Suzuki, N., Sturm, J., & Fukada, A., (2016). Survey of the views on foreign language pronunciation: Comparing students' and teachers' beliefs. Paper presented at *the 8th Pronunciation in Second Language Learning and Teaching (PSLLT) Conference*. Calgary, Canada.

Suzuki, N., Miyamoto, M., & Fukada, A. (2016).「アニメーションテロップを使用したオンライン復唱練習が発音習得に与える効果：アクセントと特殊拍を中心に」. Paper presented at *the 2016 Symposium on Japanese Language Education (AJE)*. Venice, Italy.

Miyamoto, M., & Suzuki, N. (2016). 「日本語発音指導の必要性に関する調査－学習者調査と教師調査の比較から見えてくるもの－」. Paper presented at *American Association of Teachers of Japanese 2016 Annual Spring Conference*. Seattle, WA.

Suzuki, N., & Miyamoto, M. (2016). Effects of online repetition practice with animated visual aid on the acquisition of Japanese pitch accent and special moras. Paper presented at *the 1st Purdue Languages and Cultures Conference*. Lafayette, IN.

Miyamoto, M., Mishima, H., & Yanagisawa, S. (2015). 「スピーキングを重視したオンライン日本語コースの設計・開発・運用」. Poster presented at *The 22nd Princeton Japanese Pedagogy Forum*. Princeton, NJ.

Mishima, H., Miyamoto, M., & Yanagisawa, S. (2015). 「スピーキングを重視したオンライン日本語コースの設計・開発・運用」. Paper presented at *American Association of Teachers of Japanese 2015 Annual Spring Conference*. Chicago, IL.

Miyamoto, M. (2014). 「日本語初級学習者を対象とした、オンライン口頭練習による日本語ピッチアクセントの習得効果とその練習方法」. Paper presented at *American Association of Teachers of Japanese 2014 Annual Spring Conference*. Philadelphia, PA.

GRANTS AND AWARDS

External Awards

2018	Nominated for Midwestern Association of Graduate Schools, Excellence in Teaching Award

Purdue University

2018	Certificate of Outstanding Achievement in Undergraduate Teaching
2016	School of Languages and Cultures Excellence in Teaching Award 2015-2016
2016	Office of Provost, Excellence in Distance Learning Innovation Award, $1000
2015	School of Languages and Cultures Travel Grants $300
2014	Purdue Graduate Student Travel Grants $500
2013	Graduate Student Teaching Excellence Award

COURSES TAUGHT

INSTRUCTOR/COORDINATOR, *Purdue University*

JPNS 402, Fourth Year Japanese VIII	Spring	2017
JPNS 302, Third Year Japanese VI	Spring	2014
JPNS 301, Third Year Japanese VI	Fall	2018
JPNS 241, Introduction to Japanese Literature	Fall	2018
JPNS 241, Introduction to Japanese Literature	Fall	2017

JPNS 241, Introduction to Japanese Literature	Fall	2013
JPNS 201, Second Year Japanese III, Online Course	Fall	2015
JPNS 102, First Year Japanese II	Spring	2015
JPNS 102, First Year Japanese II, Online Course	Spring	2015
JPNS 101, First Year Japanese I	Fall	2014
JPNS 101, First Year Japanese I	Summer	2012

TEACHING ASSISTANT, *Purdue University*

JPNS 202, Second Year Japanese IV	Spring	2013
JPNS 201, Second Year Japanese III	Fall	2012
JPNS 102, First Year Japanese II	Spring	2012
JPNS 101, First Year Japanese I	Fall	2017
JPNS 101, First Year Japanese I	Fall	2011

THE CULTURE AND LANGUAGE TEACHER, *Youth For Understanding USA*

Intensive culture and language program for high school students leaving Japan for study abroad.

Summer 2015
Summer 2013

RELEVANT COURSEWORK

Japanese/SLA Pedagogy Courses

Teaching Japanese
Teaching Japanese Literature
Theories in Japanese Language Acquisition
Intermediate/Advanced level Japanese Language Pedagogy
Material Development in Language Teaching
Teaching Modern Japanese Novels
Teaching Modern Japanese Popular Literature and Culture
Teaching ESL: Principles and Practices
Teaching ESL: Theoretical Foundations

Second Language Acquisition Courses

Vocabulary and Reading in Second Language Acquisition
Individual Difference in Second Language Acquisition
Acquisition of Second Language Phonology

Linguistics Courses

Introduction to English and General Linguistics
Japanese Linguistics
World Englishes
Sociolinguistics

Language Testing Courses

Seminar in Language Testing
Introduction to Measurement and Evaluation
Embedded Assessment in Language Courses

Technology Related Courses

Technological Literacy for Foreign Language Teachers
Introduction to Multi-Media Programming for Foreign Language Teaching
Computational Methods in Applied Linguistics

Research Related Courses

Quantitative Research
Experimental Methods
Introduction to Educational Research
Experimental Statistics I
Quantitative Data Analysis II
Writer's Workshop: How to Publish a Journal Article

SERVICE TO THE PROFESSION

CONFERENCE ORGANIZING COMMITTEES:

2018	Abstract Reviewer, Midwest Association of Language Testers Committee
2016-2018	Graduate Student Representative, Midwest Association of Language Testers Committee
2017	Abstract Reviewer, The 2nd Purdue Languages and Cultures Conference
2016-2017	Program Chair, The 2nd Purdue Languages and Cultures Conference Organizing Committee
2016	Graduate Student Volunteer, 18th Midwest Association of Language Testers Conference
2016	Graduate Student Volunteer, The 1st Purdue Languages and Cultures Conference
2016	Abstract Reviewer, The 1st Purdue Languages and Cultures Conference
2014	Organizing Committee member, The 14th Annual SLC Graduate Student Symposium

OTHERS:

2017 – Present	Webmaster, SLC Graduate Student Committee
2016 – Present	AATJ's National Japanese Exam (NJE) Test Development Team member
2016 – 2017	Colloquium organizing committee, School of Languages and Cultures department
2016 – 2017	Workshop coordinator, SLC Graduate Student Committee
2016 – 2017	Teaching Awards Organizing Committee
2014 – Present	Founder/Director, Japanese Cultural Events Organization at Purdue University

PROFESSIONAL EXPERIENCES

2017 June – July	Interned at Cambridge Michigan Language Assessments (CaMLA), Ann Arbor, (http://cambridgemichigan.org/blog/2017/09/21/an-interns-words-mayu-miyamoto/)
2017 – Present	Serving as a Data Analyst for National Japanese Exam (NJE) Test
2016 – Present	Member of National Japanese Exam (NJE) Test Development Team
2016 – Present	Director/Organizer of Japanese Skit & Speech Contest at Purdue University

Professional Memberships

2016 – Present	Midwest Association of Language Testers (MwALT)
2016 – Present	Washington Association for Language Teachers
2015 – Present	ヨーロパ日本語教師会 (Association of Japanese Language Teachers in Europe)
2015 – Present	日本語教育学会 (The Society for Teaching Japanese as a Foreign Language)
2014 – Present	American Association of Teachers of Japanese (AATJ)
2014 – Present	American Council on the Teaching of Foreign Languages (ACTFL)

Certificates

2016	Completed ACTFL 4-Day OPI Assessment Workshop
2014	Graduate Certificate in Teaching English as a Second Language (TESOL)
2011	Certificate in Teaching Japanese at middle and high schools in Japan
2011	Certificate in Teaching English at middle and high schools in Japan

Computer Skills

SAS for Statistical Computing
R
SPSS
JMetrik
PRAAT
Movie Maker & Camtasia
Adobe (Dream Weaver, Audition, Photoshop)
HTML, CSS, PHP
Microsoft Office (Word, PowerPoints, Excel)

Language Skills

Japanese: Native speaker
English: Near-native Speaker
Chinese: Proficient (in reading)
French: Intermediate